DOCUMENT RESUME

ED 161 918 TM 007 686

AUTHOR	Williams, Stephen R.; Folsom, Ralph E., Jr.
TITLE	Bias Resulting from Nonresponse: Methodology and Findings, Revised. Technical Report on NLS Base-Year Estimates.
INSTITUTION	Research Triangle Inst., Durham, N.C.
SPONS AGENCY	National Center for Education Statistics (DHEW), Washington, D.C.
PUB DATE	Sep 76
NOTE	83p.; For related document, see ED 133 334
EDRS PRICE	MF-$0.83 HC-$4.67 Plus Postage.
DESCRIPTORS	*Attrition (Research Studies); Followup Studies; Graduate Surveys; High School Graduates; Item Analysis; *Longitudinal Studies; *National Surveys; Questionnaires; Research Design; *Research Problems; Response Style (Tests); Sampling; School Demography; Senior High Schools; *Statistical Bias; *Statistical Data
IDENTIFIERS	*National Longitudinal Study High School Class 1972

ABSTRACT

As in any very large survey, the sample that provided data for the National Longitudinal Study of the High School Class of 1972 (NLS) differed from the random sample of the study design because of nonresponse from 226 of the 1,200 schools in the primary sample, and a small amount of other missing data. Part A of this report, briefly describes the design of the sample, including provisions for selecting alternate schools for those that declined to participate, the detailed stratification plan, and an overview of the project. Part B presents a synopsis of the findings which indicated that the majority of the estimated totals for the base year survey were significantly biased by the lack of data from the nonresponding schools. This conclusion is based on data from the first follow-up survey of 1973 which reduced the nonresponse rate to a minimal two percent. Part C describes the methodology and assumptions of the statistical procedures used in estimating the bias resulting from nonresponse. Part D tabulates the statistical results of this study. The appendix provides an alternative methodology for estimating bias in the base year statistics. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED161918

BIAS RESULTING FROM SCHOOL NONRESPONSE: METHODOLOGY AND FINDINGS

Revised

Technical Report on NLS Base-Year Estimates

Stephen R. Williams
and
Ralph E. Folsom, Jr.

Prepared for
The National Center for Education Statistics

September 1976

TM007 686

RESEARCH TRIANGLE PARK, NORTH CAROLINA 27709

TABLE OF CONTENTS

		Page
LIST OF TABLES		iii
A.	INTRODUCTION	1
	1. Sample Design	3
	2. Study Overview	4
B.	FINDINGS AND CONCLUSIONS: A SYNOPSIS	8
	1. Findings	8
	2. Conclusion	17
C.	METHODOLOGY	20
	1. Overview	21
	2. Assumptions	22
	3. Resurvey Results and Their Use in This Analysis	22
	4. Estimation of Bias in BY Statistics: Totals	25
	5. Estimation of Bias in By Statistics: Proportions	30
	6. Other Techniques of Accounting for School NR	32
	a. Unlimited Substitution	32
	b. Subgroup Weight Adjustments	33
	c. Aggregate Adjustment	34
D.	STATISTICAL RESULTS	36
REFERENCES		59
APPENDIX:	ALTERNATIVE METHODOLOGY FOR ESTIMATING BIAS IN THE BASE YEAR STATISTICS	60

LIST OF TABLES

		Page
1.	Frequency of Significant Biases in BY Estimates Related to 35 Selected Questions on the NLS Student Questionnaire	9
2.	Frequency of Significant Biases in BY Estimates and Frequency of Biases that Might Be Erroneously Declared Significant (α error).	11
3.	Bias and Reliability for BY Estimates Related to BSYRQ2	13
4.	Bias and Reliability for BY Estimates Related to BSYRQ5	16
5-39.	Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse .	37 - 58

A. INTRODUCTION

In 1969, the National Center for Educational Statistics (NCES) began planning for a national longitudinal survey of high school graduates as the first in a series of longitudinal studies of educational effects. Basic to the planning was the selection of a representative sample of students to be traced through their post-secondary education or training and then followed for some time after their entrance into the job market. The present volume relates to some of the initial results of this survey, specifically, the base year (BY) estimates for the National Longitudinal Study of the High School Class of 1972 (NLS). Especially, this volume addresses estimation biases that apparently have resulted from efforts to account for those schools in the sample that did not cooperate in the BY survey.

The primary purpose of NLS is to discover what happens to young people after they leave high school (as measured by their subsequent educational and vocational activities, plans, aspirations, and attitudes) and to relate this information to their prior educational experiences and personal and biographical characteristics. Ultimately, the study will allow a better understanding of the development of students as they pass through the American educational system and of the complex factors associated with individual educational and career outcomes. Such information is essential as a basis for effective planning, implementation, and evaluation of Federal policies and programs that are designed to enhance educational opportunity and achievement and to upgrade occupational attainment and career outcome.

A pilot phase of the NLS study was conducted in 1971 by Research Triangle Institute. This phase involved the development and field test of instruments and procedures for the BY survey. The sampling of schools was designed and executed by WESTAT Incorporated. WESTAT also developed the initial weights needed to account for unequal sampling rates (these weights are termed the unadjusted school weights). Then, in January 1972, Educational Testing Service and NCES jointly undertook the full-scale BY survey; it was completed June 30, 1973. This effort involved the completion of several instruments. The major instruments consist of student, counselor, and school questionnaires, student test booklets, and student's school record information forms. The first follow-up

data collection effort (FFU) was begun in October 1973 and was completed in April 1974. The instruments used in the FFU survey are termed a Form A Student Questionnaire for most students responding in the BY and a Form B Student Questionnaire for students from schools that did not participate in the base year. Form B was also administered to a subsample of 500 students who responded in the BY survey. At the time of this writing, the second follow-up data collection effort, which began in October 1974, has been completed and is presently being summarized.

The student questionnaire data are the major concern of the present investigation, but the contexts of all survey instruments listed above are briefly described here to point up the scope of NLS. The test booklets contained short (five to ten minute) tests of student ability in six subject areas: (1) vocabulary, (2) reading, (3) mathematics, (4) letter groups [inductive reasoning], (5) mosaic comparisons [perception], and (6) picture number [association memory]. The Student Questionnaire, which was to be completed by each sample student, consisted of 104 questions about myriad student-related factors; such as, ability, socioeconomic status, aspirations, values, religion, ethnicity, and high school experiences. Counselors, on the counselor questionnaire, were queried about their characteristics; workloads; and counseling experiences, practices, and facilities. School records and, if needed, student interviews were used to complete questions on the student's school record information form (SRIF). The SRIF contained questions about the sample students; such as, name, address, class rank, grade average, standardized test scores, transfer status, and major curriculum. The school questionnaires were usually completed with the assistance of the school's administrative staff and dealt with information about enrollment, staff, and educational practices, for example. In the FFU, Form A was administered to sample students that participated in the BY survey, and contained 85 questions about post-graduation activities and aspirations. Form B, which was mailed to students that did not participate in BY, contained the same 85 questions and, additionally, 14 questions about information that was initially sought in the BY survey. This additional information obtained from students in "nonrespondent" schools is particularly instrumental in the present investigation of bias in BY statistics.

1. Sample Design

The sample design may be described as a deeply stratified two-stage probability sample with schools as first-stage sampling units and students (also guidance counselors) as second-stage units. The target population consists of all 1972 twelfth graders enrolled in all public, private, and church-affiliated high schools in the 50 States and the District of Columbia. The first-stage sampling frame was constructed from computerized school files maintained by the United States Office of Education and by the National Catholic Educational Association.

The school sampling frame was stratified into 289 major strata based on the following variables:

- Type of control (public or nonpublic),
- Geographic region (Northeast, North Central, South, and West),
- Grade 12 enrollment (less than 300; 300 to 599; 600 or more),
- Proximity to institutions of higher learning (three categories),
- Percentage minority group enrollment (eight levels), and
- Income level of the community (11 categories for public schools, 8 categories for Catholic schools, and a single category for other schools).

Then, on the basis of the degree of urbanization, one or more final strata were defined within each major stratum to generate a total of 600 final strata.

In order to increase the numbers of disadvantaged students in the sample, schools located in low-income areas and schools with high proportions of minority group enrollments were sampled at approximately twice the sampling rate used for the remaining schools. Schools in the smallest-enrollment strata (less than 300 seniors) were selected with probabilities proportional to their estimated numbers of senior students and without replacement. Schools in the remaining strata were selected with equal probabilities and without replacement. Within each final stratum, four schools were selected initially, and then two of the four were randomly designated as the primary selections. The other two schools were retained as backup or substitute selections and were used in the sample only if one or both of the primary schools did not cooperate (for example, closed, refused, or ineligible; ineligible relates

here to schools for handicapped or legally confined students or that did not enroll students of their own). Five strata comprised exceptions to this methodology because they contained only three schools each. Samples of 18 students and two counselors per school were selected; five additional students and one additional counselor were selected as alternates. Students and counselors were selected from each sample school with equal probabilities and without replacement.

Thus, the primary sample consisted initially of 1,200 schools (two per stratum), as many as 2,400 counselors (two per sample school), and as many as 21,600 students (18 per sample school). For schools, however, the sample has ultimately involved secondary schools selected in place of primary schools that did not cooperate or had no eligible seniors; tertiary schools used to replace secondary schools that did not cooperate or had no eligible seniors; augmentation schools, 16 schools in strata 601-608 used to account for incompleteness in the original sampling frame; the occurrence of "extra" schools or schools in excess of the intended sample of two per stratum; and, in the FFU, "resurvey schools" or noncooperating BY sample schools (largely primary schools), which were surveyed during the FFU to obtain both current and retrospective data. In the BY survey, 16,409 students from 974 sample schools completed the Student Questionnaire (974 = 921 primary schools + 53 backup schools and excluded the 18 "extra" schools). In the FFU survey, 21,350 students from 1,300 sample schools completed FFU questionnaires (1,300 = 1,153 primary schools + 131 backup schools + 16 augmentation schools and excludes the 18 "extra" schools).

2. Study Overview

Extimates based on the Student Questionnaire data of the NLS were influenced by nonresponse at several levels of sample selection; for example, 230 of the initially selected schools declined to participate; approximately one-tenth of the sample students in cooperating schools failed to participate; and, finally occasional item responses were missing for participating students (median nonresponse for individual items was 2 percent [ref. 5]). This paper presents methodology and results of an investigation of the possible influence of school nonresponse on the NLS initial Base Year (BY) estimates. Overall or net bias (from all sources) and the extent to which the class-adjusted weights domiciled

on the Public Use Data File [ref. 1] account for bias were not addressed in this investigation; that is, although these topics appear worthy of consideration, they were not within the scope-of-work for this study. School nonresponse relates here to those primary schools that either refused to participate in the BY survey or could not participate because the request was received too late.* Summary findings and conclusions from this investigation are also presented in the FFU final report. Two methodologies for estimating bias, which were developed expressly for this analysis, are presented in Section C and the appendix. The basic statistics resulting from the method described in Section C are presented in Section D and consist of the following estimates:

- The number of seniors, estimated using the currently accepted number of 1972 seniors for the sample schools and the substitution and weighting methodology that were used in BY estimation (using data from 1,605 schools; 949 primary, 95 backup, and 21 with no seniors);
- The number of seniors that would respond to each category in each of 35 questions on the Student Questionnaire, estimated according to the substitution and weighting methodology and the student-response data that were used in BY estimation (using data from 1,065 schools and 16,409 students; 16,409 Student Questionnaires were completed in the 1,065 schools with 91 of these schools having no completed Student Questionnaires);
- The proportion of seniors that would respond to each of these question categories, calculated using BY methodology and data throughout (using data from 1,065 schools and 16,409 students);
- The bias in each of the above BY estimates that resulted from substituting or otherwise accounting for nonparticipating schools (using data from 1,175 schools and 18,696 students to calculate the "best estimate" and from 1,065 schools and 16,409 students to calculate the base year estimates; 1,175 schools = 949 base year primaries, 21 primaries with no seniors, and 205 primaries from FFU; 18,696 seniors - 16,409 base year respondents, plus 3,144 from FFU primary schools, minus 857 from backup schools);
- Standard deviations for each of these statistics (using the same data sets used to estimate bias).

*Noting the substantially higher response rates in follow-up activities, one might conjecture that inadequate lead time was the major cause of the school nonresponse in the BY survey.

These estimates are presented for each response category in each of thirty-five questions, FFU Form B questions 78 and 86-99 (several are multiple-part questions). Note that the Form B designations serve only to identify the questions that were investigated in this analysis. Note also that the retrospective data obtained in the FFU survey are reflected in the "best estimates," but were not used to calculate base year estimates that are described, above, as "calculated using BY methodology and data throughout."

The primary sources of information used in this investigation consist of the NLS BY survey, 1972-73, and the NLS FFU survey, 1973-74. Two components of the FFU survey, which are particularly useful in this analysis, relate to: (1) a subsample of 500 BY respondent students, who were asked to recall answers to the 35 questions listed above, and (2) a complete follow-up of the nonparticipating BY schools (BY information was sought from a sample of the 1972 seniors in these nonparticipating schools). Data used in this analysis, except for the indications of recall-bias obtained from the 500 subsampled students, were abstracted from the so-called "master file" tape from which the Public Use Data File [1] was prepared.

As noted above, approximately 20 percent of the initial-sample (primary) schools declined to participate in the 1972 NLS survey. This magnitude of nonreponse presents the potential for a substantial bias in the NLS BY estimates.

Two common procedures for dealing with the problems of nonresponse involve: (1) the selection and use of substitute sample units, and (2) the adjustment of response weights. Both of these procedures were used in the calculation of BY statistics. Each stratum was assigned two primary schools and two backup schools, which were to be used if the primary schools declined to participate. If the combined solicitation of primary and backup schools failed to yield two cooperating schools in a stratum, weight adjustments, as opposed to the solicitation of additional backup schools, were used to account for the missing schools. These procedures, as they were used in the BY survey, relied on the use of substitutes that were similar in size, geographic location, and proximate population density to the nonrespondents, and the increase of weights of participating schools that were likewise similar to nonparticipating schools.

In preview to the findings of the next section, the original BY estimates relating to each primary school are compared in this investigation with a so-called "best estimate." This comparison, or difference, constitutes the basis for estimating the school nonresponse bias of the BY estimates. The original BY estimates for a particular primary school may be based on data from the BY response of that primary school, on the BY response of a substitute (backup) school, or, implicity, on other BY responses through weight adjustments. The "best estimate" is based on the currently most reliable data about the primary schools; for most NR schools in the BY, these data have been obtained retrospectively from the resurvey activity. The "best estimates" used in this investigation do not, however, utilize certain recent refinements, such as, the weighting-class adjustments that were used to account for NR in the FFU estimates and the 16 augmentation schools that were used to account for sampling-frame incompleteness. These refinements, if used to obtain these "best estimates," would result more nearly in estimates of the overall bias as opposed to that resulting only from the school nonresponse, which we attempted to isolate in this particular investigation.

B. FINDINGS AND CONCLUSIONS: A SYNOPSIS

1. Findings

The results of this study should be viewed with the awareness that several simplifying assumptions and approximations were invoked. While the accuracy of the inference statements to follow depend to some degree on the validity of these assumptions and approximations, the results are so consistent and dramatic it is doubtful that a more refined analysis would alter the conclusions. On the other hand, considerable effort in this investigation has been directed at minimizing errors. BY statistics were recalculated and verified against previous results, and a test run of the computer program produced correct results for all items, according to manual calculations, using actual data from more than 50 schools.

The findings herein are couched largely in the form of statistical inferences, recognizing that differences were sure to exist between an initial sample of "primary" schools and a backup sample of substitute schools, but that such differences might be largely attributable to sampling error. For each category of each question, the null hypothesis (H_o) of negligible bias was tested at three significance levels, namely, α = .10, .05, and .01 with two sided standard normal critical regions. Table 1 presents a summary of these tests. The question-identification (BSYRQ) numbers correspond to the original base-year questionnaire. The table reveals that the vast majority of the estimated totals are significantly biased--mostly negatively. More noteworthy, however, is the predominance of biases for estimated proportions because the BY statistics were presented as proportions or ratio estimates. Considering the large number of hypotheses being tested (several categories for each of the 35 questions), one should expect to reject H_o: BIAS = 0 a certain number of times even if the difference were attributable only to sampling error. It is useful, therefore, to compare the number of categories in which H_o is rejected versus the number of rejections that could be expected to result from the commission of type I(α) errors. Table 2 presents this comparison and reveals that H_o was rejected far too often to be accounted for entirely by α error. For the .05 significance

Table 1. Frequency of Significant Biases in BY Estimates Related to 35 Selected Questions on the NLS Student Questionnaire*

BSYRQ Number	Question	Number of categories	Frequency of categories with significant bias**			
			Estimated totals		Estimated proportions	
			$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$
2	(high school program)	8	0	3	1	3
5	(high school grades)	8	1	6	1	4
8	(job time)	8	1	7	0	0
10A	(athletics)	3	0	3	0	0
10B	(cheerleader)	3	0	2	0	1
10C	(debating or music)	3	0	2	1	0
10D	(hobby clubs)	3	0	1	0	0
10E	(honorary clubs)	3	1	1	0	0
10F	(school newspaper or yearbook)	3	1	2	0	0
10G	(subject matter clubs)	3	0	1	0	0
10H	(student government)	3	0	2	0	1
10I	(vocational education clubs)	3	0	2	0	0
16	(friends plans)	8	0	4	0	1
27	(time of decision on college)	6	0	4	0	3
83	(physical limitation)	2	0	1	0	0
84	(race)	8	1	1	0	4
88	(English spoken at home)	2	1	1	0	2
90A	(father's education)	9	1	6	0	4
90B	(mother's education)	9	0	8	0	3
91A	(father's aspirations for student's education)	7	0	5	2	0
91B	(mother's aspirations for student's education)	7	0	5	0	0
92	(religion)	6	0	4	1	0
93	(parent income)	10	0	0	0	0
94A	(place for home study)	2	0	2	0	0
94B	(daily newspaper)	2	0	2	0	0
94C	(dictionary)	2	0	1	0	0
94D	(encyclopedia)	2	0	1	0	0
94E	(magazines)	2	1	1	0	0

Table 1. (continued)

94F	(record player)	2	0	1	0	0
94G	(tape recorder)	2	0	2	0	0
94H	(color television)	2	0	2	0	0
94I	(typewriter)	2	0	2	0	0
94J	(electric dishwasher)	2	0	2	0	0
94K	(motor vehicles)	2	0	2	0	0
95	(type of community)	8	0	2	1	0
	Total	155	8	91	7	26

*Based on data from the NLS BY, FFU, and Recall-Error surveys, 1972 and 1973; each rejected category is listed only once under the most powerful test that it passes.

**The test statistic, $\frac{\hat{\text{Bias}}}{\text{SE}_{\hat{\beta}}}$, is assumed to have approximately the standard normal distribution.

Table 2. Frequency of Significant Biases in BY Estimates and Frequency of Biases that Might Be Erroneously Declared Significant (α Error)*

Type of Estimator	Significance Level, α	Biases Declared Significant, Number	Expected Number of Type I Errors**
Totals	0.10	99	16
Proportions	0.10	33	16
Totals	0.05	91	8
Proportions	0.05	26	8

*Based on data from the NLS BY, FFU, and Recall-Error Surveys, 1972 and 1973.

**Expected number of times H_o: $\beta = 0$ might be rejected because Type I (α) error is calculated as α times the total number of tests (155).

level, for example, we would expect to reject approximately eight times each for totals and ratios if H_o were true, but, rather, we rejected H_o 91 and 26 times, respectively.

The estimated number of 1972 seniors, based on BY methodology and the presently accepted numbers of 1972 seniors in sample schools, is indicated by this study to contain a statistically-significant downward bias of five percent. This bias suggests that the larger schools were underrepresented as a result of substitution and weight-adjustment actions taken in the BY to compensate for school NR.

Biases and their implications can be analysed for each question individually from the statistical results presented in Section D. Such individual-question analyses were not undertaken here. An example is presented, however, to demonstrate pertinent considerations regarding the biases and their magnitudes as they relate to the accuracy of BY estimation. The first two BY questions that were studied, BSYRQ2 and BSYRQ5, are used for this example, but these two questions are not particularly unique compared to the biases indicated throughout the analysis.

BSYRQ2 relates to the student's high school program. The question and its eight answer-categories are:

Which of the following best describes your present high school program?

(Circle one.)

General. 1
Academic or college preparatory. 2
Vocational or technical:
  Agricultural occupations 3
  Business or office occupations 4
  Distributive education 5
  Health occupations 6
  Home economics occupations 7
  Trade or industrial occupations. 8

Recall from table 1 that three categories for estimated totals and four for estimated proportions were significantly biased for this question. Table 3 reveals that, for totals, categories 1, 2, and 4 are negatively biased. The relative biases of these category totals range in magnitude from a minus 3.4 percent for the number answering "Academic" to a minus

Table 3. Bias and Reliability for BY Estimates Related to BSYRQ2[a]

Question Category	Type of Estimate	"Best Estimate"[b]	BY Estimate[c]	Relative Bias in BY Estimate, Percent	$\frac{\sqrt{MSE}}{SE}$ for BY Estimate	Confidence Level for the $\hat{\theta} \pm 1.96$ SE Interval
		(1)	(2)	(3)	(4)	(5)
All Categories	total	3,016,526	2,862,893	-5.1***	4.2	0.02
1	total	1,019,685	921,217	-9.7***	5.3	0.00
2	total	1,265,639	1,223,047	-3.4**	1.9	0.63
3	total	45,149	38,826	-14.0	2.3	0.46
4	total	357,950	324,546	-9.3**	3.2	0.15
5	total	72,833	81,686	12.2	2.0	0.58
6	total	24,790	25,318	2.1	1.0	0.94
7	total	30,016	28,235	-5.9	1.2	0.91
8	total	163,581	152,905	-6.5	1.8	0.67
1	proportion	.3422	.3294	-3.9**	2.5	0.36
2	proportion	.4242	.4343	2.3*	2.0	0.60
3	proportion	.0152	.0144	-5.6	1.3	0.87
4	proportion	.1202	.1161	-3.5	1.6	0.77
5	proportion	.0245	.0295	16.9***	3.3	0.12
6	proportion	.0084	.0101	16.8**	2.6	0.32
7	proportion	.0103	.0108	4.6	1.1	0.91
8	proportion	.0549	.0554	0.9	1.0	0.94

*, **, *** Significant at the .10, .05, and .01 percent levels, respectively, with two-sided standard normal critical regions.

a) Based on data from the NLS BY, FFU, and Recall-Error Surveys, 1972 and 1973.

b) Termed "best estimate" for purposes of assessing the biasing effect of school non-response comparison in this study and does not actually comprise the best estimate available; for example, it does not reflect adjustments for sampling frame incompleteness or weighting class adjustments that can be used to compensate for student NR in cooperating schools. Also, category totals were not forced to sum to what might be termed one of the more reliable estimators of number of seniors in all categories, one based on student record information.

c) These BY estimates were calculated specifically for this study using BY methodology and data that were available at that time, but because of recent revisions to the number of seniors in some sample schools, these BY estimates differ slightly from previously published BY estimates.

9.7 percent for the number of students indicating "General" programs. For ratio estimates, three categories showed significant positive bias and one showed significant negative bias. One of the more noteworthy observations here relates to the significance of bias in the ratio estimators. Three categories for ratio estimators are significant at the .05 level or higher. These categories relate to the proportion estimated in the BY for "General" school program, which is indicated to be negatively biased by 3.9 percent, the proportion in "Distributive education," which is positively biased by 16.9 percent, and the proportion in 'Health" with a positive bias of 16.8 percent. One might conjecture that the overrepresented small schools have proportionately more "Distributive and Health Education" students and fewer "General Education" students as compared to the larger schools.

A measure of the importance of bias can be obtained by comparing it to sampling error. One such comparison, $\sqrt{MSE}/SE$, is presented in table 3. MSE becomes more pertinent than the standard error when a significant bias influences the accuracy of a statistic; MSE can be represented by the following equation:

$$MSE = SE^2 + BIAS^2$$

where

MSE $\equiv$ mean square error and

SE $\equiv$ sampling error.

When the ratio, $\sqrt{MSE}/SE$, is near unity, the bias is relatively unimportant, and the standard deviation or sampling error (SE) may be used to describe the accuracy of an estimate. Note that for the proportion estimates presented in table 3, only two, those for categories 7 and 8, may be satisfactorily described using SE (or coefficient of variation as was used to report BY statistics). The proportion estimates for categories 3 and 4 may also be satisfactorily described by SE because the bias estimates, although larger, are relatively imprecise. When this ratio, $\sqrt{MSE}/SE$, is less than approximately 1.4, the MSE can be used to relate estimation accuracy and can be interpreted, without serious error, as if it were the SE. In other words, with $\sqrt{MSE}/SE \leq 1.4$, the confidence interval of $\pm \sqrt{MSE}\, Z(1-\alpha/2)$ can, in practice, be interpreted in the same way as if the estimate were unbiased and $\pm$ SE $Z(1-\alpha/2)$

were used to set two sided (1-α)-level confidence intervals where Z(1-α/2) is the upper 1-α/2 percentage point of the standard normal distribution [3]. For larger values of this ratio, such as those indicated for the proportion estimates for categories 1, 2, 4, 5, and 6, the reliability of the estimate can be represented by the area of the standard normal distribution between the values

$$\frac{BIAS}{SE} - Z(1-\alpha/2) \text{ and } \frac{BIAS}{SE} + Z(1-\alpha/2).$$

As can be seen in the final column of table 3, accuracy of the proportion estimates with relatively large √MSE/SE ratio was seriously overstated by using only SE. For the proportion estimated for category 1, for example, the use of SE alone would result in the statement that the true proportion was within ± 1.96 SE of the estimated proportion with a confidence level of .95 when, in fact, this statement holds with a confidence level of only .36.

BSYRQ5, the second question studied, reflects substantially the same severity of school NR bias. This question relates to high school grades:

Which of the following best describes your grades so far in high school?

(Circle one.)

Mostly A (a numerical average of 90-100). 1
About half A and half B (85-89) 2
Mostly B (80-84). 3
About half B and half C (75-79) 4
Mostly C (70-74). 5
About half C and half D (65-69) 6
Mostly D (60-64). 7
Mostly below D (below 60) 8

The analysis of this question suggests, as pointed up in table 4, that the overrepresentation of the relatively smaller schools, which was indicated to result from substitution and weight adjustment for school NR, produced a substantial downward bias in the estimated proportion of high school students receiving above average grades and an equally significant upward bias in proportion receiving low grades. For this question also, the final column of table 4 reveals that the accuracy of proportion estimates is substantially overstated in all but two of the eight question-categories.

Table 4. Bias and Reliability for BY Estimates Related to BSYRQ5[a]

Question category	Type of estimate	"Best estimate"[b]	BY estimate[c]	Relative bias in BY estimate, percent	$\frac{\sqrt{MSE}}{SE}$ for BY estimate	Confidence level for the $\hat{\theta} \pm 1.96$ SE interval
		(1)	(2)	(3)	(4)	(5)
All categories	total	3,016,526	2,862,893	-5.1***	4.2	0.02
1	total	286,803	276,994	-3.4*	1.3	0.85
2	total	597,034	532,879	-10.7***	4.5	0.01
3	total	629,319	569,670	-9.5***	5.5	0.00
4	total	804,900	781,085	-3.0**	1.9	0.63
5	total	423,608	399,017	-5.8**	2.5	0.35
6	total	206,052	189,534	-8.0**	2.4	0.42
7	total	26,046	30,005	15.2**	2.1	0.54
8	total	5,910	7,367	24.7	1.7	0.74
1	proportion	.0966	.0994	2.8	1.3	0.85
2	proportion	.2002	.1913	-4.7**	2.6	0.33
3	proportion	.2108	.2043	-3.2**	2.2	0.50
4	proportion	.2703	.2807	3.7**	2.8	0.26
5	proportion	.1422	.1429	0.5	1.0	0.94
6	proportion	.0692	.0679	-1.9	1.1	0.91
7	proportion	.0088	.0107	17.8***	2.9	0.23
8	proportion	.0022	.0028	21.4***	1.8	0.68

*, **, *** significant at the .10, .05, and .01 percent levels, respectively, with two-sided standard normal critical regions.

(a) Based on data from the NLS BY, FFU, and Recall-Error surveys, 1972 and 1973.

(b) Termed "best estimate" for purposes of assessing the biasing effect of school nonresponse comparison in this study and does not actually comprise the best estimate available; for example, it does not reflect adjustments for sampling frame incompleteness or weighting class adjustments that can be used to compensate for student NR in cooperating schools. Also, category totals were not forced to sum to what might be termed one of the more reliable estimators of number of seniors in all categories, one based on student record information.

(c) These BY estimates were calculated specifically for this study using BY methodology and data that were available at that time, but because of recent revisions to the number of seniors in some sample schools, these BY estimates differ slightly from previously published BY estimates.

In order to utilize a consistent set of data on number of seniors enrolled at each school, the most recent and reportedly most accurate series was utilized. This resulted in estimated totals being slightly lower than those originally published (approximately 2 percent), but this difference should not have had any significant influence on estimated proportions or estimated biases for either totals or proportions. The appropriate way to adjust BY estimates, if such adjustments are made on the basis of these bias estimates, is to adjust the published estimates of proportions as opposed to adjusting proportions estimated in this study. To adjust for the indicated bias in the proportion estimated for category 1 of BSYRQ2, for example, the -3.9 percent should be added to the published 32.87 [4] rather than to the 32.94 presented in table 3. The adjusted percentages could be calculated according to the following equation:

$$\hat{P} = \hat{P}(B)/(\hat{\beta} + 1)$$

where

$\hat{P}$ ≡ the estimated percent of students responding to a particular category adjusted for school NR bias,

$\hat{P}(B)$ ≡ the corresponding biased BY estimate, and

$\hat{\beta}$ ≡ the estimated relative bias expressed as a proportion of $\hat{P}$.

In the example of category 1 of BSYRQ2, the adjusted percentage would equal 32.9/(-.039) + 1) or 34.2. This example reflects the fact that, where totals or proportions as previously published differ from those in the present study, the bias that is attributable to school NR is more reliably indicated by the bias estimate (as in column 3) than by the difference between the published estimate and the "best estimate" calculated in this study.

2. Conclusion

This analysis of possible bias in BY estimates, which relates to 35 questions on the NLS Student Questionnaire, indicates that school non-response (NR) substantially affected many of the statistics developed from the BY survey. The NR effect, which will be termed NR bias, was identified by viewing the difference between BY estimates as they were calculated versus estimates that incorporate the additional information obtained during the First Follow-Up survey of 1973. This difference necessarily relates almost exclusively to school NR bias because the difference between the

two estimates results solely from reducing the school NR from approximately 20 percent to a minimal two percent.

The average NR bias is approximately a negative five percent for estimated totals, such as the estimated number of eligible seniors in 1972, or the estimated number of seniors who would respond in a particular question-category on the NLS student questionnaire. This difference suggests that, within a major stratum, the larger schools declined to participate more frequently than did the smaller schools. Note that irrespective of whether substitute schools or weight adjustments were used to account for school NR, the responses and characteristics of one or more cooperating schools within that major stratum were used in place of the NR schools.

The BY methodology also involved the substitution of positively responding schools for schools that had no eligible seniors, were closed, or did not exist--"valid zeros." Admittedly, such a substitution is not appropriate for the estimation of totals, but the effect was apparently small--resulting in approximately a one percent increase in estimated totals. Estimates of totals presented in this report reflect the zeros as valid responses.

Ratio statistics (proportions) have been used thus far in presenting the NSL findings, and, providing that student-response proportions are approximately the same for larger as opposed to smaller schools, the negative bias in totals would not be of particular concern. The analysis indicates, however, that the response proportions are not uniform by size of school and that substantial bias from school NR is also reflected in the BY estimates of ratios or proportions. These biases are alternately positive and negative for each question because if one question-category is biased in one direction then one or more of the remaining categories will reflect this bias in the opposite direction. Statistically significant bias was indicated for estimated proportions in 14 of the 35 questions studied. Approximately one-third of the questions had one or more categories that reflected proportion-estimate bias significant at the 5 percent level or higher.

In view of these findings, several topics become pertinent for consideration. These include:

- Because of this bias, reliability statements for many of the BY ratio estimates were probably overstated - users of BY statistics should be cognizant of this.

- When comparing followup statistics against the BY statistics, allowances should be made for the bias in those proportions in which bias was indicated. If patterns can be identified in the types of responses that are upward biased, downward biased, or not biased, rough approximations of bias might also be inferred for selected questions that were not investigated in this study. This possibility and the extent to which this bias is accounted for by the class-adjusted weights on the Public Use Data File [ref. 1], however, have not been dealt with here.
- Initially, RTI proposed to investigate the individual effectiveness of several techniques that might be used to deal with the problem of school NR. The computer software for this additional work has largely been developed, but has not yet been utilized. In view of the implications of the findings thus far, however, it now seems that this additional analysis may be instructive, whereas, if no significant bias had been found in the BY ratios, the additional analysis would have been somewhat academic.
- When class comparisons (age-sex specific, for example) are contemplated, valid interpretation of observed differences may be precluded by biases of the magnitude indicated unless these biases are accounted for. Therefore, although biases at the subclass level were not estimated in this investigation, their usefulness should also be considered.
- The study findings clearly demonstrate that the use of backup schools (and probably backup students) cannot be relied upon to produce unbiased estimates. Hence, in future surveys, greater emphases might advisably be placed on obtaining higher response rates from the initial sample, with less or no reliance being placed on the use of the backups.

C. METHODOLOGY

In this section, the methodology is presented that was used to obtain the estimates reported in Section D, below, and in the NLS final report. The two methodologies developed (see the appendix also) in the study invoke simplifying although different assumptions, and both rely on Taylor linearizations of nonlinear statistics as a basis for estimating variance. It was not ascertained which of the two methodologies is superior, so that it should be noted that the methodology described in this section was used primarily because it involved somewhat less software development.

1. Overview

In previous papers on NLS methodology, [2] for example, the estimation of bias was proposed using equations of the form:

$$\hat{\beta} = \hat{X}(B) - \hat{X} \qquad (1.1)$$

where

$\hat{\beta} \equiv$ the estimated bias of BY totals that resulted from school nonresponse;

$\hat{X}(B) \equiv$ the BY total, biased by school nonresponse, and is a weighted sum of the following type

$$\hat{X}(B) = \sum_{h=1}^{600} \sum_{i=1}^{m_h} \sum_{j=1}^{n_{hi}} W_{hij} X_{hij}$$

$$= \sum_{h=1}^{600} \sum_{i=1}^{m_h} \hat{X}_{hi}$$

$$= \sum_{h=1}^{600} \hat{X}_h \; ; \quad \text{and}$$

$\hat{X} \equiv$ the BY total corrected for school-nonresponse bias based on a follow-up of noncooperating primary schools; and subscripts h, i, and j relate to stratum, school, and student, respectively.

$\hat{X}(B)$ and $\hat{X}$ are two estimates of totals, such as, the number of seniors that would indicate a specific question-category on the Student Questionnaire. These two estimates differ only in the values of weights, W_{hij}, and the values of student responses, X_{hij}, for those primary schools that did not participate

in the BY survey. In this analysis, two (m_h=2) school observations are used in each estimation model from each of the 600 final strata, with n_{hi} student observations per sample school. The weights, $W_{hij} = W_{uhij}\ [N_{hi}/n_{hi}]$, are student weights adjusted for nonresponse within participating schools by letting n_{hi} depict the count of responding seniors in place of the number selected. The X_{hij} are taken to be 1 or 0, depending upon whether a response category is selected or not selected by student hij. Notice that $W_{hij} = W_{hi}$ is constant for each of the j=1 (1) n_{hi} students. Also, W_{uhi} in the calculation of $\hat{X}$ is the original BY unadjusted school weight for primary school hi.

As demonstrated in Section 2, below, equation 1.1, which appropriately expresses the bias as a difference between two interdependent estimators, can be simplified. That is, the $\hat{\beta}$ of equation 1.1 can be recast as a weighted sum of the type used to estimate $\hat{X}$ and $\hat{X}(B)$, thus facilitating the use of the same estimation and variance equations as those already being used in NLS analyses. Otherwise stated, $\hat{X}(B)_{hi}$, the "half-stratum" total based on school hi, is calculated for each primary school according to whether the school responded in the BY; was substituted for in the BY; or was estimated in the BY on the basis of responses from similar schools (weight adjustments). The difference between $\hat{X}(B)_{hi}$ and the $\hat{X}_{hi}$, which is corrected by resurvey information, comprises the bias indication for primary school hi expanded up to represent a "half stratum" bias for school nonresponse in that stratum. In this difference, or $\hat{\beta}_{hi}$ form, the $\hat{\beta}$ have straightforward variance estimates.

Thus, the bias for school nonresponse is reduced to the linear statistic

$$\hat{\beta} = \hat{X}(B) - \hat{X}$$

$$= \sum_{h=1}^{600} \sum_{i=1}^{2} \hat{\beta}_{hi} \qquad (1.2)$$

with estimated variance

$$V(\hat{\beta}) = \sum_{h=1}^{600} [\hat{\beta}_{h1} - \hat{\beta}_{h2}]^2 .$$

Variances for ratio statistics and their biases are approximated using Taylor linearizations. The details of these linearizations are presented, below, in Section 5.

2. Assumptions

In each stratum, two "difference" or "bias" values are observed, one for each primary-sample school. These values are the differences between student information later obtained for that school and the information actually used in the 1972 BY estimates; often these two values are identical resulting in a "difference" or "bias" value of zero. Each of these "bias" indications is taken to be an independent observation from the population of schools in that particular stratum (an observation that presumably reflects the propensity to cooperate and related factors for the schools in that stratum). In fact, however, each of the observations is not precisely an independent observation because its value will depend upon which school was selected as the other sample school for the stratum, which sample schools were selected as substitute schools, and which schools were selected in other final strata in the particular major strata, as well as the willingness of each of these schools to participate and their characteristics.

By taking the expected value of $\hat{\beta}$ in equation 1.1, and noting that $\hat{X}$ is a relatively unbiased (in terms of school nonresponse bias) estimate of X for a specific question-category, one may note that $\hat{\beta}$ is, by definition, an unbiased estimate of the bias in $\hat{X}(B)$. Variances of $\hat{\beta}$, however, may be slightly understated by this methodology because the covariance terms, $\text{cov}[\hat{\beta}_{h1}, \hat{\beta}_{h2}]$, which are assumed to be zero in the calculation of stratum variances, are suspected to have small positive values. The $\hat{\beta}_{hi}$ quantities alluded to are the expanded-up bias measure for schools i = 1 and 2 presented in the previous section.

3. Resurvey Results and Their Use in This Analysis

As noted in Section A, above, a FFU subsample of 500 students was asked to recall their answers to selected questions that had been asked on the BY Student Questionnaire. This information was used in the present investigation to account for possible memory bias in BY information collected retrospectively in the FFU survey and, hence, in the calculation of $\hat{X}$ in equation 1.1. The selected BY student data that were obtained retrospectively in the 1973 FFU survey involved a sample of 1972 seniors of the primary-sample schools that did not participate in the BY survey. Because this information about school NR bias was obtained retrospectively, it is

itself biased by recall errors. To evaluate and account for this recall bias, the 500 randomly-selected respondent students (students who had responded in the BY survey) were asked to recall 1972 information for 15 questions on the BY Student Questionnaire. The questions, numbers 78 and 86-99 on the FFU Form B, actually involve a total of 35 questions (some are multiple-part questions).

For each of the 35 questions, a k x k matrix of n_{fg} values were calculated, where

$k-1 \equiv$ the number of alternative responses allowed for a particular question (row k and column k present frequencies of improper responses, such as, refusals, multiple responses, out-of-range responses, and "other" or "Don't know" when these are not explicit options), and

$n_{fg} \equiv$ the number of sample students giving response f in the FFU survey and response g in the BY survey.

Several statistical tests were made on each question to ascertain its relative stability (amount of recall bias). One of these tests, for example, indicated that one-half (17) of the questions can be reported with 80 percent consistency, that is about 80 percent of the students would enter the same answer in 1973 as they did in 1972. The statistical test, which is relied upon for this analysis (NR bias), views the marginal totals ($n_{.g}$ and $n_{f.}$) of these matrices as being (k-1) dimensional multinomial variates with vector parameters of $\underline{\pi}_1$ and $\underline{\pi}_2$, respectively. The hypotheses that were tested include:

H_o: $\underline{\pi}_1 = \underline{\pi}_2$

H_a: $\pi_{1f} \neq \pi_{2g}$ for at least one f (f = g = 1, ..., k-1).

Because, in the present methodology, the unacceptable responses are distributed proportionately, this category is not of particular interest in adjusting for NR bias. The test of homogeneity that excluded the proportion of unacceptable responses, therefore, was used in this analysis as the test criteria and indicated that the following 16 questions contained significant recall bias:

78 A, B	92 A, B
87	93
88	94 A, H, J
89 A, H	96
90	97 .
91	

FFU responses to these questions were adjusted for use in this study according to the following equation to arrive at a "best estimate" of BY response for primary schools that did not respond in the BY survey. For a particular question and category, say category g, we have:

$$\hat{X}_{ghi} = \hat{X}^{(AFF)}_{ghi} R_{hi} \qquad g = 1, \ldots, k-1 \tag{3.1}$$

where

$\hat{X}_{ghi} \equiv$ the "best estimate" for the category g of a particular question for school hi,

$\hat{X}^{(AFF)}_{ghi} \equiv$ the First Follow-Up response, $\hat{X}^{(FF)}_{ghi}$, corrected for memory or recall bias (see below), and

$R_{hi} \equiv$ a factor that distributes or "smears" the unacceptable answers over the remaining k-1 categories for that question (see below).

Further, $$\hat{X}^{(AFF)}_{ghi} = \sum_{f=1}^{k} n_{fg} \hat{X}^{(FF)}_{fhi} / n_{f.} \tag{3.2}$$

where the $\hat{X}^{(\ )}_{ghi}$ are as defined for 3.1, and

$n_{fg} \equiv$ the number of students (of the the 500) that reported category f in the FFU and category g in the BY survey, and

$$n_{f.} = \sum_{g=1}^{k} n_{fg} .$$

Finally, $$R_{hi} = \sum_{g=1}^{k} \hat{X}^{(FF)}_{ghi} \Big/ \sum_{g=1}^{k-1} \hat{X}^{(AFF)}_{ghi} .$$

Responses to the remaining 19 questions were used without adjustment for recall bias; that is,

$$\sum_{f=1}^{k} \frac{n_{fg}}{n_{f.}} \hat{X}_{fhi}^{(FF)} = \hat{X}_{ghi}^{(AFF)} = \hat{X}_{ghi}^{(FF)}$$

and

$$\hat{X}_{ghi} = \hat{X}_{ghi}^{(FF)} R_{hi}$$

for all k-1 categories. The unacceptable responses were "smeared" for all questions.

4. Estimation of Bias in BY Statistics: Totals

The methodology developed here gains some of its appeal from the fact that the statistic $\hat{\beta}$ is an unbiased estimator of the bias in $\hat{X}(B)$; that is,

$$E(\hat{\beta}) = E[\hat{X}(B) - \hat{X}]$$

$$= E[\hat{X}(B)] - E[\hat{X}]$$

$$\doteq X + BIAS_{\hat{X}(B)} - X$$

$$\doteq BIAS_{\hat{X}(B)},$$

and can be reduced to

$$\sum_{h=1}^{600} \sum_{i=1}^{2} \left[\hat{X}(B)_{hi} - \hat{X}_{hi}\right], \text{ or}$$

$$\sum_{h=1}^{600} \sum_{i=1}^{2} \hat{\beta}_{hi}.$$

Proof:

From 1.1 we have:

$$\hat{\beta} = \hat{X}(B) - \hat{X}$$

$$= \sum_{h=1}^{600} \sum_{i=1}^{2} \sum_{j=1}^{n_{hi}} W_{hij} X_{hij} - \sum_{h=1}^{600} \sum_{i=1}^{2} \sum_{j=1}^{n'_{hi}} W'_{hij} X'_{hij}$$

$$= \sum_{h=1}^{600} \sum_{i=1}^{2} \left[\sum_{j=1}^{n_{hi}} W_{hij} X_{hij} - \sum_{j=1}^{n'_{hi}} W'_{hij} X'_{hij}\right]$$

with the prime (') indicating primary school values, which may or may not be the same as the unprimed counterpart. Notice that the two terms in the brackets are "half-stratum" expansions for school hi for $\hat{X}(B)$ and $\hat{X}$, respectively.

Continuing with the proof,

$$\hat{\beta} = \sum_{h=1}^{600} \sum_{i=1}^{2} [\hat{X}(B)_{hi} - \hat{X}_{hi}] ,$$

$$= \sum_{h=1}^{600} \sum_{i=1}^{2} \hat{\beta}_{hi} \text{ as was to be shown.}$$

where

$\hat{X}(B)_{hi} \equiv$ either an expanded response for a substitute school, the estimated value of the companion school (primary or substitute) in that final stratum, or an average estimated value for the major stratum ℓ that contains stratum h,

$\hat{X}_{hi} \equiv$ a "best estimate" for primary school hi (expanded to the half-stratum level),

$W'_{hij} \equiv$ the student weight, adjusted for student nonresponse, for the primary school, or

$$W'_{hij} = [A'_h/2A'_{hi}][N'_{hi}/n'_{hi}] = W'_{hi}$$

with

$N'_{hi} \equiv$ the number of senior students in primary school hi,

$n'_{hi} \equiv$ the number of sample seniors that finally responded on the Student Questionnaire either in BY or FFU,

$$A'_{hi} \equiv \begin{cases} 1 & \text{for all primary schools in the large-school (300+) strata;} \\ \text{Presurvey estimate of senior enrollment for primary schools in the small-school (<300) strata.} \end{cases}$$

$$A'_h = \sum_{i=1}^{M_h} A'_{hi} \equiv \text{total measure for all schools in stratum h in the entire sampling frame.}$$

The biased estimate, $\hat{X}_{hi}(B)$, takes several forms depending on the BY response experience in stratum h. Consider the following possibilities:

(1) Primary school hi participated in the BY survey,

then $\hat{X}(B)_{hi} = \hat{X}_{hi}$ and $\hat{\beta}_{hi} = 0$. (4.1)

(2) Primary school hi declined to participate in the BY survey and a substitute response was used in its stead, then

$$\hat{X}(B)_{hi} = \hat{X}_{hs} = W_{hs} X_{hs}$$

where the subscript s denotes a substitute school and

$$X_{hs} = \sum_{j=1}^{n_{hs}} X_{hsj}$$

with

$$X_{hsj} = \begin{cases} 1 & \text{if student j in the substitute school indicated the particular question-category,} \\ 0 & \text{otherwise} \end{cases}$$

and

$$\hat{\beta}_{hi} = \hat{X}_{hs} - \hat{X}_{hi} \quad . \tag{4.2}$$

(3) Primary school hi declined to participate in the BY survey and only one school in stratum h participated, this school being either the companion primary or a substitute for the companion primary. The weight of the cooperating school is doubled to account for this nonresponse. Thus,

$$\hat{X}(B)_{hi} = \begin{cases} \hat{X}_{hi'}, & \text{if primary school hi' participated;} \\ \hat{X}_{hs} & \text{if a substitute school s participated;} \end{cases}$$

and correspondingly,

$$\hat{\beta}_{hi} = \begin{cases} \hat{X}_{hi'} - \hat{X}_{hi} \\ \text{or} \\ \hat{X}_{hs} - \hat{X}_{hi} \quad . \end{cases} \tag{4.3}$$

(4) Primary school hi in major stratum ℓ along with its companion primary and both substitutes in stratum h declined to participate. Then the weights of all participating schools in stratum ℓ are inflated by the factors $C_{\ell h}$,

where:

$$C_{\ell h} = \{1 + A_{o\ell} / A_{\ell h} \; H_{\ell}\} \; , \tag{4.4}$$

with the ℓh subscript combination denoting final stratum h within major stratum ℓ, and

$A_{o\ell} \equiv$ the accumulation of size measures, $A(\ell h)$, for all final strata in ℓ that had no participating schools in BY,

$H_\ell \equiv$ the number of final strata in ℓ with at least one participating school.

Adjusting weights of all BY participating schools in major stratum ℓ by the smearing factor $C_{\ell h}$ is equivalent to substituting for each primary school in a noncooperating stratum ℓh the contribution

$$\hat{X}(B)_{hi} = A_h \bar{\bar{X}}_\ell / 2$$

where $$\bar{\bar{X}}_\ell = \sum_{h \in BY(\ell)} \left[\sum_{i=1}^{m_h} W_{ahi} \hat{\hat{X}}_{hi} \Big/ \sum_{i=1}^{m_h} W_{ahi} A_{hi} \Big/ H_\ell \right]$$

and A_{hi} = the presurvey size measure for school hi,

$\sum_{h \in BY(\ell)} \equiv$ summation over all final stratum h in major stratum ℓ that have at least one (m_h=1) participating school;

$\hat{\hat{X}}_{hi} \equiv$ the school level estimate, $N_{hi} X_{hi}/n_{hi}$, and

$W_{ahi} \equiv$ the adjusted school weight; that is

$$W_{ahi} = 2W_{uhi}/m_h = A_h/m_h A_{hi} .$$

Thus, $\hat{\beta}_{hi} = A_h \bar{\bar{X}}_\ell / 2 - \hat{\hat{X}}_{hi}$ and we notice that $\bar{\bar{X}}_\ell$ for small-enrollment schools is a proportion and for large-enrollment schools is a school average. This result can be verified by noting that (4.5)

$$\hat{X}(B)_\ell - \hat{X}_\ell = \sum_{h \in BY(\ell)} \sum_{i=1}^{m_h} W^*_{ahi} \hat{\hat{X}}_{hi} - \sum_{h \in BY(\ell)} \sum_{i=1}^{2} W'_{uhi} \hat{\hat{X}}'_{hi} - \sum_{h \notin BY(\ell)} \sum_{i=1}^{2} W'_{uhi} \hat{\hat{X}}'_{hi} \qquad (4.6)$$

where the prime (') denotes observations and weights for a primary school, $W^*_{ahi} = C_{\ell h} W_{ahi}$, and other terms are according to definitions for 4.5.

Notice that $\hat{X}(B)_\ell$ can be expanded to:

$$\sum_{h\in BY(\ell)} \sum_{i=1}^{m_h} W^*_{ahi} \hat{X}_{hi} = \sum_{h\in BY(\ell)} \sum_{i=1}^{m_h} W_{ahi} \hat{X}_{hi}$$

$$+ \left\{ \sum_{h\notin BY(\ell)} \sum_{i=1}^{2} W_{uhi} A_{hi} \right\} \left\{ \left[\sum_{h\in BY(\ell)} \sum_{i=1}^{m_h} W_{ahi} \hat{X}_{hi} \Big/ \sum_{i=1}^{m_h} W_{ahi} A_{hi} \right] / H_\ell \right\}$$

$$= \sum_{h\in BY(\ell)} \sum_{i=1}^{m_h} W_{ahi} \hat{X}_{hi} + \sum_{h\notin BY(\ell)} \sum_{i=1}^{2} W_{uhi} A_{hi} \bar{\bar{X}}_\ell \qquad (4.7)$$

so, for strata $h\notin BY(\ell)$, from 4.6 and 4.7,

$$\hat{X}(B) - \hat{X} = \sum_{h\notin BY(\ell)} \sum_{i=1}^{2} W_{uhi} (A_{hi} \bar{\bar{X}}_\ell - \hat{X}'_{hi}) .$$

Thus, for each school in $h\notin BY(\ell)$

$$X(B)_{hi} - \hat{X}_{hi} = A_h \bar{\bar{X}}_\ell / 2 - \hat{X}_{hi}$$

as was to be shown.

The variance for $\hat{\beta}$ can be estimated on the basis of the variation of these $\hat{\beta}_{hi}$ values, two in each stratum. As noted earlier, this assumes that each $\hat{\beta}_{hi}$ is an independent observation from stratum h, an assumption that may cause $\text{var}(\hat{\beta})$ to be slightly understated. Also, note that the clusters of final stage sampling units (students) in this characterization are not defined in advance and become clustered or associated with one another through a complex conditional selection process that depends on which schools are selected for substitutes and whether they cooperate when needed. Irrespective of the complexities of this clustering process, cluster totals will contain contributions from both the "between" and "within" cluster variation, and can be used to approximate the pertinent variances.

The following equations are proposed for the estimation of the variance of $\hat{\beta}$:

$$\widehat{\text{Var}}(\hat{B}) = \sum_{h\in R} m_h s^2_{h\beta} + \sum_{h\notin R} m_h s^2_{h'\beta} \qquad (4.8)$$

where $\sum_{h \in R}$ is the summation over those strata in which a valid estimate of variance of $\hat{\beta}_{hi}$ is calculable; $h \in R$ being the compliment strata. Valid bias or $\hat{\beta}_{hi}$ values were not calculable for 26 (2 percent of the sample) schools on the basis of resurvey efforts. For these 26, data from other schools with comparable nonresponse patterns were used to estimate $\hat{\beta}_{hi}$ and the second term of 4.8, above.

And,

$$s^2_{h\beta} = \sum_{i=1}^{m_h} [\hat{\beta}_{hi} - \bar{\beta}_h]^2/(m_h - 1)$$

$$= [\hat{\beta}_{hi} - \hat{\beta}_{h2}]^2/2 \qquad \text{when } m_h = 2 .$$

5. Estimation of Bias in BY Statistics: Proportions

For each question-category that was treated in the resurvey, the bias $\hat{\beta}$ was estimated for the BY proportions or averages. These estimators, $\hat{\bar{X}}$, are the ratio estimates because the population and subpopulation totals of the denominator are not generally known. The numbers of senior students for most schools are known, approximately, but in general the numbers of students in subgroups must be estimated. The bias for these ratio statistics can be estimated according to the equations in Section 4 above, for the numerator and denominator individually, but the bias of the ratio is not the ratio of the biases, rather:

$$\hat{\beta}_R = \hat{\bar{X}}(B) - \hat{\bar{X}}$$

$$= \frac{\hat{X}(B)}{\hat{N}(B)} - \frac{\hat{X}}{\hat{N}}$$

$$= \frac{\hat{X} + \hat{\beta}_x}{\hat{N} + \hat{\beta}_n} - \frac{\hat{X}}{\hat{N}} \qquad (5.1)$$

Equation 5.1 can be quantified by simply substituting the prior estimates of totals. The variance of this composite statistic cannot be calculated as before (Section 4), but is approximated by $\sum_h d_h^2$ where d_h is derived below using a Taylor linearization:

$$F = \frac{\sum_h \sum_i \hat{X}_{hi} + \sum_h \sum_i \hat{\beta}_{xhi}}{\sum_h \sum_i \hat{N}_{hi} + \sum_h \sum_i \hat{\beta}_{Nhi}} - \frac{\sum_h \sum_i \hat{X}_{hi}}{\sum_h \sum_i \hat{N}_{hi}}$$

where

$\hat{X}$ and $\hat{X}(B)$ are according to equations defined in Section 4,

$\hat{N}(B)$ ≡ the summation of $\hat{X}(B)$ for all categories in a particular question or an estimate of the total number of seniors,

$\hat{\beta}_x$ and $\hat{\beta}_N$ ≡ bias estimates developed in the previous section with the subscripts denoting estimated category total and estimated number of seniors, respectively, and

h and i ≡ subscripts designating strata 1 to 600 and schools 1 and 2 within each stratum.

The following are partial derivatives with respect to each sample draw variate, evaluated at expected values.

$$\left.\frac{\partial F}{\partial \hat{X}_{hi}}\right|_E = \frac{1}{N+\beta_N} - \frac{1}{N}$$

$$\left.\frac{\partial F}{\partial \hat{\beta}_{xhi}}\right|_E = \frac{1}{N+\beta_N}$$

$$\left.\frac{\partial F}{\partial \hat{\beta}_{Nhi}}\right|_E = \frac{-(X+\beta_x)}{(N+\beta_N)^2}$$

$$\left.\frac{\partial F}{\partial \hat{N}_{hi}}\right|_E = -\frac{(X + \beta_x)}{(N + \beta_N)^2} + \frac{X}{N^2}$$

$$\text{And } F' = \sum_h \sum_i \left\{ \left(\frac{1}{N+\beta_N} - \frac{1}{N} \right) \hat{X}_{hi} + \frac{\hat{\beta}_{xhi}}{N+\beta_N} - \frac{(X + \beta_x)}{(N + \beta_N)^2} \hat{\beta}_{Nhi} \right.$$

$$\left. - \frac{(X + \beta_x)}{(N + \beta_N)^2} \hat{N}_{hi} + \frac{X}{N^2} \hat{N}_{hi} \right\}.$$

$$\text{So, } \hat{Var}(F) \doteq \sum_h \left\{ \left(\frac{1}{N + \beta_N} - \frac{1}{N} \right) \left(\hat{X}_{h1} - \hat{X}_{h2} \right) + \frac{\hat{\beta}_{xh1} - \hat{\beta}_{xh2}}{N + \beta_N} \right.$$

$$- \frac{(X + \beta_x)}{(N + \beta_N)^2} (\hat{\beta}_{Nh1} - \hat{\beta}_{Nh2}) - \frac{(X + \beta_x)}{(N + \beta_N)^2} (\hat{N}_{h1} - \hat{N}_{h2})$$

$$\left. + \frac{X}{N^2} (\hat{N}_{h1} - \hat{N}_{h2}) \right\}^2$$

$$= \sum_h d_h^2 \quad \text{referred to, above.}$$

b. Other Techniques of Accounting For School NR

The methodology used to investigate school NR bias in the BY estimates is extended here to address other selected techniques of accounting for school NR. These techniques are named 1) unlimited substitution, 2) subgroup weight adjustments, and 3) aggregate adjustment. The computer software for these investigations, which correspond to topics suggested by Moore [2], was only partially completed in the present effort.

a. Unlimited Substitution

An indication is sought here of the bias that would result from unlimited use of substitute schools within each final stratum; that is, the number of contacts would not be limited to four schools as they were in the BY survey. The following bias estimator is proposed to evaluate an estimated total for a particular question-category.

$$\hat{\beta}_1 = \sum_h \left\{ \sum_{i \in BY} (\hat{X}(1)_{hi} - \hat{X}_{hi}) + \sum_{i \notin BY} (\hat{X}(1)_{(hi)'} - \hat{X}_{(hi)'}) \right\} \qquad (6.1)$$

where

The range of summation BY relates to all primary schools that either participated or were substituted for in the BY survey;

The (hi)' subscript pertains to the use of the previous indication of substitution bias for those primary schools that were imputed in the BY by adjusting weights and, hence, do not constitute a valid indication of substitution bias;

$\hat{\beta}_1$ ≡ the estimated bias for the statistic, $\hat{X}(1)$, being investigated in this section;

$\hat{X}(1)_{hi}$ ≡ the biased estimator using substitution only for stratum h and primary school i (see (4.2); and

$\hat{X}_{hi}$ ≡ the "best estimate" for school hi (see 3.1).

Variance of $\hat{\beta}_1$ is according to 4.8 except for the summation ranges:

$$\hat{Var}(\hat{\beta}_1) = \sum_{h \in BY} \sum_{i=1}^{2} (\hat{\beta}_{1h1} - \hat{\beta}_{1h2})^2$$

$$+ \sum_{h \notin BY} \sum_{i=1}^{2} (\hat{\beta}_{1(h1)'} - \hat{\beta}_{1(h2)'})^2 \qquad (6.2)$$

where terms are defined as in 6.1.

b. Subgroup Weight Adjustments

Here, the methodology is presented to investigate the amount of bias that might result with no school substitution, but, rather, from using weight adjustments at the lowest possible level. This method of accounting for school NR suggests that nonrespondents in a final stratum h (or major stratum ℓ) are more characteristic of respondents in that stratum than respondents from some other stratum. The following bias estimator is proposed for evaluating school NR bias in estimated totals using these weighting techniques.

$$\hat{\beta}_2 = \sum_h \left\{ \sum_{i \in FF} (\hat{X}(2)_{hi} - \hat{X}_{hi}) + \sum_{i \notin FF} (\hat{X}(2)_{(hi)'} - \hat{X}_{(hi)'}) \right\} \qquad (6.3)$$

where:

The range of summation FF relates to all primary schools except the 26 that have declined to participate in both BY and FFU surveys;

The (hi)' subscript pertains to the use of the previous indication of weight-adjustment bias for the 26 noncooperating schools alluded to above;

$\hat{X}_{hi} \equiv$ the "best estimate" for school hi (see (3.1);

$\hat{\beta}_2 \equiv$ the estimated bias for the statistic being investigated in this section; and

$$\hat{X}(2)_{hi} = \begin{cases} \hat{X}_{hi} \text{ if the primary participated in the BY} \\ \\ \text{According to equations 4.3 or 4.5 otherwise.} \end{cases}$$

Variance of $\hat{\beta}_2$ is according to 6.2 except the summation ranges are changed to correspond to $\hat{\beta}_2$.

c. Aggregate Adjustment

Here, a methodology is presented to evaluate bias that can be expected to result with a "minimum treatment" of NR, the use only of a single, aggregate adjustment. The following bias estimator, $\hat{\beta}_3$, relates to a particular question-category and to an estimated total.

$$\hat{\beta}_3 = \sum_h \left\{ \sum_{i \in FF} (\hat{X}(3)_{hi} - \hat{X}_{hi}) + \sum_{i \notin FF} (\hat{X}(3)_{(hi)'} - \hat{X}_{(hi)'}) \right\}$$

where

The range of summation and the (hi)' subscript are as defined for 6.3;

$\hat{\beta}_3 \equiv$ the estimated bias for the statistic, $\hat{X}(3)$, being investigated in this section; and

$\hat{X}_{hi} \equiv$ the "best estimate" for school hi (see 3.1).

$$\hat{X}(3)_{hi} = \begin{cases} \hat{X}_{hi} \text{ if school hi participated in the BY, and} \\ N_{hi} \sum_{h} \sum_{i \varepsilon BY} W_{uhi} \hat{X}_{hi} \Big/ \sum_{h} \sum_{i \varepsilon BY} W_{uhi} N_{hi} \cdot \end{cases}$$

The summation range, BY, pertains to all primaries responding in the BY; N_{hi} is the recorded number of seniors in school hi; and the variance of $\hat{\beta}_3$ is according to 6.2 except the summation ranges are changed to coincide with $\hat{\beta}_3$.

D. STATISTICAL RESULTS

The tabular results of this section relate to two estimator forms, the estimated number of seniors that would respond to each category of 35 questions, and the proportions of seniors that would respond in each of these categories. The methodology used to estimate the biases of these BY statistics is presented in Section C above. For proportion statistics, this bias, or net influence of accounting for nonresponding schools, relates to statistics as they were calculated in the base year; note, however, that the statistics were actually recalculated for use in this analysis, so that any alterations in base year weights or data tapes would have minimal effect on the estimation of bias (steps were taken throughout to ensure that the bias estimates would, as nearly as possible, relate solely to school nonresponse).

The bias for totals, on the other hand, relates to statistics that are based on the BY methodology, only as it related to school substitution and weight adjustment. Only proportions were presented in the summary statistics of the NLS BY reports, so that the pertinent form of a statistic for totals was necessarily somewhat arbitrary. The results should, however, provide a useful indication of bias that might occur in totals when using techniques such as those used in the BY to account for school nonresponse, irrespective of the exact estimator form. The statistic used to estimate totals in the following results is the usual summation of "expanded-up" responses that is used to obtain the numerator and denominator values of the NLS proportion statistics; except, "zero schools" are included as valid responses. In the BY methodology, these "zero schools" were substituted for. A "zero school" is one that had no eligible 1972 seniors, was closed, or did not exist at the time of the BY survey.

Table 5. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ2. Which of the following best describes your present high school program?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
General	921	-98***	32.94	-1.28**
	(19)	(21)	(0.55)	(0.51)
Academic	1,223	-43**	43.43	1.01*
	(26)	(19)	(0.59)	(0.56)
Agricultural	39	-6	1.44	-0.08
	(3)	(6)	(0.10	(0.15)
Business	325	-33**	11.61	-0.41
	(11)	(16)	(0.34)	(0.52)
Distributive Education	82	9	2.95	0.50***
	(5)	(6)	(0.16)	(0.16)
Health	25	1	1.01	0.17**
	(2)	(3)	(0.07)	(0.08)
Home Economics	28	-2	1.08	0.05
	(3)	(4)	(0.09)	(0.09)
Trade	153	-11	5.54	0.05
	(7)	(8)	(0.23)	(0.24)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 6. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ5. Which of the following best describes your grades so far in high school?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Mostly A	277	-10*	9.94	0.28
	(11)	(6)	(0.31)	(0.24)
Half A and half B	533	-64***	19.13	-0.89**
	(15)	(15)	(0.37)	(0.37)
Mostly B	570	-60***	20.43	-0.65*
	(11)	(14)	(0.33)	(0.36)
Half B and half C	781	-24**	28.07	1.04**
	(15)	(12)	(0.40)	(0.43)
Mostly C	399	-25**	14.29	0.07
	(11)	(12)	(0.31)	(0.26)
Half C and half D	190	-17**	6.79	-0.13
	(8)	(7)	(0.23)	(0.17)
Mostly D	30	4**	1.07	0.19***
	(2)	(2)	(0.07)	(0.04)
Mostly below D	7	1	0.28	0.06***
	(1)	(2)	(0.04)	(0.02)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 7. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ8. On the average over the school year, how many hours per week do you work in a paid or unpaid job? (Exclude vacations.)

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
None	674	-50***	24.31	-0.04
	(13)	(10)	(0.37)	(0.42)
Less than 6	320	-24***	11.51	-0.04
	(8)	(8)	(0.24)	(0.24)
6 to 10	342	-30**	12.30	-0.19
	(10)	(12)	(0.26)	(0.31)
11 to 15	274	-20**	9.82	-0.03
	(7)	(8)	(0.22)	(0.25)
16 to 20	381	-21*	13.67	0.18
	(10)	(11)	(0.26)	(0.28)
21 to 25	286	-22***	10.25	-0.08
	(8)	(4)	(0.23)	(0.20)
26 to 30	196	-11***	7.10	0.11
	(6)	(4)	(0.18)	(0.13)
More than 30	306	-21***	11.04	0.07
	(8)	(8)	(0.26)	(0.19)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 8. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ10A. Have you participated in athletic teams, intramurals, letterman's club or sports club, either in or out of school this year?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Have not participated	1,561 (25)	-100*** (26)	56.13 (0.51)	0.40 (0.73)
Have participated actively	943 (19)	-81*** (27)	33.95 (0.43)	-0.42 (0.59)
Have participated as a leader or officer	275 (8)	-19*** (6)	9.92 (0.24)	0.01 (0.24)

Table 9.

Question: BSYRQ10B. Have you participated in cheerleaders, pep club or majorettes, either in or out of school this year?

Have not participated	2,298 (30)	-133*** (37)	82.49 (0.55)	0.86 (0.86)
Have participated actively	375 (17)	-53*** (16)	13.47 (0.45)	-0.89** (0.41)
Have participated as a leader or officer	112 (5)	-8 (6)	4.04 (0.16)	0.02 (0.16)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 10. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ10C. Have you participated in debating, drama, band or chorus, either in or out of school this year?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Have not participated	1,852 (28)	-151*** (37)	66.49 (0.50)	-0.64 (0.78)
Have participated actively	738 (16)	-46*** (16)	26.59 (0.40)	0.22 (0.44)
Have participated as a leader or officer	191 (7)	-2 (5)	6.92 (0.21)	0.39* (0.21)

Table 11.

Question: BSYRQ10D. Have you participated in hobby clubs such as photography, model building, hot rod, electronics and crafts, either in or out of school this year?

Have not participated	2,256 (32)	-171*** (45)	81.13 (0.47)	-0.35 (0.98)
Have participated actively	460 (12)	-20 (13)	16.36 (0.33)	0.28 (0.32)
Have participated as a leader or officer	70 (4)	-3 (5)	2.51 (0.13)	0.05 (0.12)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 12. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ10E. Have you participated in honorary clubs such as Beta Club or National Honor Society, either in or out of school this year?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Have not participated	2,347 (33)	-167*** (44)	84.18 (0.53)	-0.14 (0.90)
Have participated actively	363 (13)	-21* (11)	12.96 (0.38)	0.06 (0.31)
Have participated as a leader or officer	78 (5)	-4 (6)	2.86 (0.16)	0.06 (0.15)

Table 13.

Question: BSYRQ10F. Have you participated in school newspaper, magazine, yearbook or annual, either in or out of school this year?

Have not participated	2,186 (31)	-153*** (41)	78.64 (0.60)	0.17 (0.85)
Have participated actively	438 (15)	-36*** (12)	15.70 (0.41)	-0.22 (0.37)
Have participated as a leader or officer	155 (8)	-12* (7)	5.66 (0.25)	0.03 (0.18)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 14. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ10G. Have you participated in school subject matter clubs such as science, history, language, business or art, either in or out of school this year?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Have not participated	2,040 (31)	-162*** (35)	73.48 (0.54)	-0.66 (0.96)
Have participated actively	625 (15)	-23 (17)	22.13 (0.41)	0.55 (0.46)
Have participated as a leader or officer	123 (6)	-6 (9)	4.39 (0.19)	0.09 (0.18)

Table 15.

Question: BSYRQ10H. Have you participated in student council, student government or political club, either in or out of school this year?

Have not participated	2,215 (33)	-182*** (43)	79.64 (0.47)	-0.78 (0.92)
Have participated actively	383 (11)	-20** (10)	13.78 (0.32)	0.23 (0.32)
Have participated as a leader or officer	181 (7)	2 (8)	6.58 (0.22)	0.53*** (0.17)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 16. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ10I. Have you participated in vocational education clubs such as Future Homemakers, Teachers, Farmers of America, DECA, OEA, FBLA, or VICA, either in or out of school this year?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Have not participated	2,138 (34)	-140*** (33)	76.67 (0.57)	0.37 (0.99)
Have participated actively	462 (13)	-51*** (19)	16.71 (0.39)	-0.61 (0.58)
Have participated as a leader or officer	181 (7)	-8 (6)	6.62 (0.21)	0.22 (0.25)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 17. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ16. What do most of your close friends plan to do next year?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Military Service	89	0	3.21	0.22
	(5)	(2)	(0.17)	(0.14)
Vocational/ trade school	209	-23**	7.62	-0.21
	(8)	(9)	(0.24)	(0.26)
Homemaker	49	-14**	1.79	-0.33**
	(4)	(6)	(0.12)	(0.13)
College	1,594	-122***	57.27	-0.22
	(30)	(25)	(0.65)	(0.58)
On-the-job training	34	4	0.93	-0.10
	(2)	(6)	(0.07)	(0.08)
Work	414	-20**	14.87	0.29
	(12)	(9)	(0.36)	(0.31)
Don't Know	303	-11	10.99	0.39
	(9)	(10)	(0.28)	(0.30)
Other	93	-8	3.32	-0.07
	(5)	(6)	(0.15)	(0.17)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 18. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ27. When did you first decide whether you would go to college or not?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Before tenth grade	1,193	-81***	42.75	0.04
	(22)	(17)	(0.51)	(0.49)
In tenth grade	207	-24***	7.45	-0.31***
	(7)	(5)	(0.20)	(0.12)
In eleventh grade	339	-36***	12.14	-0.43**
	(9)	(8)	(0.24)	(0.20)
This year	519	-37***	18.67	-0.03
	(12)	(11)	(0.31)	(0.33)
Still undecided	527	-16	18.99	0.71**
	(11)	(14)	(0.32)	(0.35)

Table 19.

Question: BSYRQ83. Do you have a physical condition that limits the kind or amount of work you can do on a job?

No	2,636	-184***	94.61	-0.03
	(37)	(47)	(0.41)	(0.97)
Yes	152	-8	5.39	0.02
	(7)	(11)	(0.22)	(0.20)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 20. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ84. How do you describe yourself?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
American Indian	34	1	1.25	0.14**
	(3)	(7)	(0.09)	(0.07)
Black, Afro-American or Negro	265	-25*	9.64	-0.29
	(13)	(15)	(0.41)	(0.48)
Mexican-American or Chicano	86	11	3.13	0.58**
	(8)	(13)	(0.27)	(0.28)
Puerto Rican	13	2	0.50	0.11**
	(2)	(13)	(0.05)	(0.05)
Other Latin-American	22	2	0.78	0.10**
	(3)	(6)	(0.09)	(0.05)
Oriental or Asian-American	27	-1	0.97	0.03
	(3)	(6)	(0.10)	(0.08)
White or Caucasian	2,267	-176***	80.93	-0.84
	(36)	(44)	(0.58)	(1.01)
Other	80	1	2.80	0.14
	(4)	(9)	(0.14)	(0.12)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 21. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ88. Is English the language spoken most often in your home?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
No	233	-128***	8.36	-3.61***
	(12)	(17)	(0.37)	(0.69)
Yes	2,539	-79*	91.64	3.59***
	(35)	(42)	(0.49)	(0.93)

Table 22.

Question: BSYRQ90A. What was the highest educational level your father or male guardian completed?

Doesn't apply	98	-8**	3.65	0.06
	(5)	(3)	(0.15)	(0.14)
Did not complete high school	813	-91***	29.46	-0.92
	(19)	(19)	(0.46)	(0.59)
High school or equivalent	824	-77***	29.78	-0.41
	(16)	(20)	(0.40)	(0.59)
Adult education program	42	-15*	1.51	-0.40**
	(5)	(9)	(0.15)	(0.16)
Business or trade school	160	17***	5.79	0.95***
	(7)	(6)	(0.24)	(0.22)
Some college	298	-20***	10.78	0.11
	(8)	(6)	(0.26)	(0.18)
Finished college	286	-32***	10.30	-0.31
	(10)	(8)	(0.29)	(0.23)
Graduate school	77	6	2.79	0.38**
	(5)	(4)	(0.14)	(0.15)
Graduate or professional level	165	3	5.94	0.51***
	(8)	(4)	(0.24)	(0.19)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 23. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ90B. What was the highest educational level your mother or female guardian completed?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Doesn't apply	71	-10**	2.56	-0.15
	(4)	(4)	(0.14)	(0.13)
Did not complete high school	702	-46***	25.28	0.14
	(16)	(14)	(0.44)	(0.53)
High school or equivalent	1,189	-107***	42.70	-0.76
	(22)	(29)	(0.47)	(0.54)
Adult education program	69	-19***	2.47	-0.46**
	(4)	(7)	(0.14)	(0.19)
Business or trade school	170	14***	6.06	0.84***
	(7)	(4)	(0.21)	(0.14)
Some college	278	-15**	9.99	0.17
	(8)	(6)	(0.25)	(0.17)
Finished college	198	-17***	7.07	-0.14
	(8)	(6)	(0.25)	(0.19)
Graduate school	50	4**	1.79	0.25***
	(3)	(2)	(0.11)	(0.07)
Graduate or professional degree	57	-2	2.08	0.08
	(4)	(2)	(0.11)	(0.07)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 24. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ91A. As far as you know, how much schooling does your father (or guardian) want you to get?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Quit high school w/o graduating	7 (1)	0 (2)	0.27 (0.04)	0.04 (0.08)
Graduate from high school	203 (7)	-6 (7)	7.47 (0.23)	0.43* (0.25)
Graduate from h/s then trade school	571 (16)	-52*** (13)	20.71 (0.43)	-0.24 (0.33)
Two-year junior college	268 (9)	-22*** (7)	9.63 (0.26)	-0.08 (0.27)
Four-year college	1,007 (19)	-75*** (20)	36.28 (0.45)	-0.03 (0.49)
Graduate or professional school	321 (11)	-35*** (10)	11.50 (0.31)	-0.41* (0.24)
Don't know	391 (10)	-21*** (8)	14.14 (0.30)	0.27 (0.21)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 25. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ91B. As far as you know, how much schooling does your mother (or guardian) want you to get?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Quit high school w/o graduating	10	1	0.34	0.04
	(1)	(2)	(0.04)	(0.06)
Graduate from high school	188	-6	6.89	0.35
	(7)	(7)	(0.21)	(0.22)
Graduate from h/s then trade school	608	-49***	22.03	-0.08
	(16)	(13)	(0.44)	(0.35)
Two-year junior college	302	-30***	10.81	-0.28
	(9)	(8)	(0.27)	(0.28)
Four-year college	1,045	-72***	37.60	0.11
	(19)	(17)	(0.46)	(0.45)
Graduate or professional school	338	-29***	12.09	-0.21
	(11)	(9)	(0.31)	(0.21)
Don't know	284	-20**	10.23	0.05
	(8)	(8)	(0.25)	(0.22)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 26. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ92. What religion were you brought up in?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Protestant	1,033	-104***	37.02	-1.18
	(19)	(28)	(0.60)	(0.74)
Roman Catholic	922	-33**	32.76	0.83
	(29)	(16)	(0.73)	(0.53)
Other Christian	494	-35***	17.79	-0.05
	(13)	(13)	(0.37)	(0.36)
Jewish	82	1	2.89	0.19*
	(8)	(5)	(0.27)	(0.11)
Other	115	-6**	4.15	0.09
	(4)	(3)	(0.14)	(0.14)
None	151	-6	5.38	0.10
	(6)	(7)	(0.20)	(.21)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 27. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ93. What is the approximate income before taxes of your parents (or guardian)? Include taxable and nontaxable income from all sources.

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category,	Bias in estimated total,	Estimated proportion of seniors in each category,	Bias in estimated proportions, percentage points
(Per Year)	thousands	thousands	percent	
Less than $3,000	166	-1	5.60	-0.09
	(6)	(29)	(0.18)	(0.71)
$ 3,000 - $ 5,999	320	-15	10.64	-0.73
	(8)	(33)	(0.27)	(0.94)
$ 6,000 - $ 7,499	319	8	10.48	0.09
	(10)	(36)	(0.29)	(1.07)
$ 7,500 - $ 8,999	299	21	9.91	0.56
	(8)	(40)	(0.25)	(1.25)
$ 9,000 - $10,499	349	-14	11.48	-0.69
	(11)	(24)	(0.29)	(0.44)
$10,500 - $11,999	295	-4	9.72	-0.31
	(9)	(37)	(0.27)	(1.11)
$12,000 - $13,499	305	13	10.00	0.27
	(8)	(47)	(0.24)	(1.52)
$13,500 - $14,999	238	28	7.80	0.84
	(7)	(47)	(0.22)	(1.50)
$15,000 - $18,000	288	11	9.42	0.22
	(10)	(35)	(0.28)	(1.05)
Over $18,000	456	7	14.95	-0.06
	(14)	(32)	(0.40)	(0.89)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 28. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ94A. Do your parents have a specific place for study in their home?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Yes	1,703	-148***	61.34	-0.79
	(27)	(29)	(0.49)	(0.78)
No	1,073	-55**	38.66	0.77
	(20)	(24)	(0.44)	(0.51)

Table 29.

Question: BSYRQ94B. Do your parents have a daily newspaper in their home?

Yes	2,483	-175***	89.41	0.31
	(36)	(45)	(0.43)	(0.99)
No	291	-31***	10.59	-0.34
	(9)	(12)	(0.30)	(0.28)

Table 30.

Question: BSYRQ94C. Do your parents have a dictionary in their home?

Yes	2,739	-192***	98.29	-0.05
	(39)	(49)	(0.37)	(1.02)
No	48	-1	1.71	0.05
	(4)	(13)	(0.12)	(0.15)

[a/] Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

[b/] Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 31. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ94D. Do your parents have an encyclopedia or other reference books in their home?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Yes	2,482	-179***	89.08	-0.18
	(37)	(44)	(0.43)	(0.99)
No	303	-16	10.92	0.17
	(8)	(13)	(0.27)	(0.26)

Table 32.

Question: BSYRQ94E. Do your parents have magazines in their home?

Yes	2,568	-173***	92.23	0.22
	(37)	(43)	(0.41)	(0.88)
No	217	-21*	7.77	-0.24
	(7)	(11)	(0.24)	(0.29)

Table 33.

Question: BSYRQ94F. Do your parents have a record player in their home?

Yes	2,671	-185***	95.86	0.05
	(38)	(46)	(0.39)	(0.97)
No	115	-9	4.14	-0.05
	(6)	(11)	(0.19)	(0.29)

[a/] Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

[b/] Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 34. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ94G. Do your parents have a tape recorder or cassette player in their home?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Yes	1,988	-141***	71.20	-0.20
	(31)	(35)	(0.48)	(0.89)
No	803	-48**	28.80	0.18
	(16)	(19)	(0.41)	(0.46)

Table 35.

Question: BSYRQ94H. Do your parents have a color television in their home?

Yes	1,735	-133***	62.87	0.29
	(28)	(31)	(0.49)	(0.81)
No	1,020	-92***	37.13	-0.32
	(19)	(22)	(0.45)	(0.66)

Table 36.

Question: BSYRQ94I. Do your parents have a typewriter in their home?

Yes	2,286	-140***	81.80	0.41
	(35)	(38)	(0.47)	(0.84)
No	508	-46***	18.20	-0.43
	(12)	(14)	(0.36)	(0.32)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 37. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ94J. Do your parents have an electric dishwasher in their home?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Yes	1,061	-80***	38.47	0.28
	(24)	(19)	(0.56)	(0.71)
No	1,682	-156***	61.53	-0.32
	(33)	(35)	(0.58)	(0.98)

Table 38.

Question: BSYRQ94K. Do your parents have two or more cars or trucks that run?

Yes	2,103	-139***	75.67	0.50
	(33)	(39)	(0.56)	(0.77)
No	674	-64***	24.33	-0.52
	(17)	(18)	(0.47)	(0.54)

a/ Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

b/ Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

Table 39. Base Year Estimates Relating to Selected Questions on the NLS Student Questionnaire: Bias Resulting from School Nonresponse[a/]

Question: BSYRQ95. When you were a senior in high school which of the following best described the location of the place in which you lived?

Response Category	Selected Statistics[b/]			
	Estimated total for seniors in each category, thousands	Bias in estimated total, thousands	Estimated proportion of seniors in each category, percent	Bias in estimated proportions, percentage points
Rural	575	-73***	21.28	-0.60
	(20)	(26)	(0.64)	(0.89)
City of 50,000 of fewer	704	-106***	25.59	-1.46
	(27)	(36)	(0.80)	(1.01)
City of 50,000 - 100,000	343	-22	12.50	0.22
	(18)	(15)	(0.55)	(0.53)
Suburb of city of 50,000 - 100,000	228	6	8.29	0.82*
	(10)	(12)	(0.31)	(0.43)
City of 100,000 - 500,000	275	-19	10.04	0.17
	(16)	(14)	(0.52)	(0.51)
Suburb of city of 100,000 - 500,000	253	-11	9.21	0.35
	(12)	(12)	(0.36)	(0.43)
City over 500,000	158	-13	5.78	0.02
	(9)	(12)	(0.31)	(0.42)
Suburb of city over 500,000	201	-4	7.31	0.43
	(11)	(11)	(0.36)	(0.41)

[a/] Based on data from the NLS Base Year and First Follow-Up surveys, and the methodology of Section C. Proportions are based on Base Year methodology throughout, but a modification in the handling of "zero" schools was incorporated to estimate totals (columns one and two), hence the estimated proportions do not coincide exactly to the estimated totals.

[b/] Standard deviations are presented in parentheses and *, **, *** indicate statistical significance of the bias estimates at the 0.10, 0.05, 0.01 α-error levels, respectively.

REFERENCES

[1] Levinsohn, J., Riccobono, J. A., and Moore, R. P., "Base-Year and First Follow-up Data File Users Manual," (RTP, N.C.: Research Triangle Institute, 1975).

[2] Moore, R. Paul, "NLS Analysis Plans, Sampling Paper Number 1," (RTP, N.C.: Research Triangle Institute, 1974).

[3] Hansen, M. H., Hurwitz, W. N., and Madow, W. G., Sampling Survey Methods and Theory, Vol. 1, (New York: Wiley and Sons, 1953), p. 57.

[4] Office of Education, NCES, "National Longitudinal Study of the High School Class of 1972: Appendix B," June 1973.

[5] Fetters, William B., "Student Questionnarie and First Followup Questionnaire Item Nonresponse Rates," September 25, 1975 memorandum.

APPENDIX

ALTERNATIVE METHODOLOGY FOR ESTIMATING BIAS IN THE BASE YEAR STATISTICS

ALTERNATIVE METHODOLOGY

1.0 NLS Base-year Survey and Follow-up Design

As noted earlier, the NLS base-year survey design was a stratified two-stage selection of schools and students within schools. Within each of the 600 NLS final strata, four schools were initially selected with two of the four randomly identified as backup schools. If either of the primary selections refused to cooperate a backup school was contacted. If necessary, both backup schools were contacted in an attempt to obtain two responding schools per stratum. During NLS's first follow-up, school data were sought for all noncooperative primary selections.

2.0 Sampling Model for NLS Base-year Survey and Follow-up

In a typical stratum, the four initial school selections were made without replacement and with probabilities strictly proportional to relative size measures

$$P(i) = \begin{cases} [1/N] \text{ for large school strata} \\ [A(i)/A(+)] \text{for remaining strata} \end{cases} \qquad (2.1)$$

where N represents the total number of schools in the stratum and A(i) denotes a presurvey estimate of the senior enrollment for school-i. Also,

$$A(+) = \sum_{i=1}^{N} A(i).$$

The subsequent analysis will proceed as if the number of schools asked to cooperate in a typical stratum, say n, were fixed. In fact, however, for with replacement school selections, one can show that n has the truncated negative binomial distribution with $R = \sum_{i \in \mathcal{R}} P(i)$ corresponding to the probability of a success where $\mathcal{R}$ represents the set of schools which would cooperate

if contacted. The number of successes (responding schools) desired is $m = 2$. Recalling the negative binomial distribution; namely

$$\Pr\{n=n|m,R\} = \binom{n-1}{m-1}R^m(1-R)^{n-m} \text{ for } n = m(1)\infty \quad . \tag{2.2}$$

With $m = 2$ and truncation after $n = 4$, one notes that

$$\Pr\{n=2\} = R^2;\ \Pr\{n=3\} = 2R^2[1-R];\ \Pr\{n=4\} = [1+2R][1-R]^2$$

which leads to the expected value and variance of n,

$$E\{n\} = 4(1-R^2 + R^3/2);\ \text{Var}(n) = 2R^2(3-R-8R^2 + 8R^3).$$

Due to the truncation, m (the number of responding schools) is also a random variable jointly distributed with n as follows:

$$\Pr\{m,n\} = \begin{cases} (1-R)^4 & \text{if } m = 0,\ n = 4 \\ 4R(1-R)^3 & \text{if } m = 1,\ n = 4 \\ R^2 & \text{if } m = 2,\ n = 2 \\ 2R^2(1-R) & \text{if } m = 2,\ n = 3 \\ 3R^2(1-R)^2 & \text{if } m = 2,\ n = 4 \\ 0 & \text{otherwise} \quad . \end{cases} \tag{2.3}$$

From the joint distribution of m and n in (2.3) one observes that

$$E(m) = 4R(1-R^2 + R^3/2) = RE(n)$$

and

$$\text{Var}(m) = 2R^2(6-8R + 3R^2) - RE(n)[RE(n) - 1].$$

While it is clear that

$$\{E(m)/E(n)\} = R \quad ,$$

the usual estimator for R, namely $\hat{R} = (m/n)$, is biased. From (2.3) one obtains

$$E(m/n) = R + (5-3R)\ R^2(1-R)/6 \quad . \tag{2.4}$$

While the analysis above indicates that there are some problems associated with the assumption that n is fixed (when in fact, it is a random variable with a

coefficient of variation of approximately 50 percent) and that m is a binomial variable with mean nR and variance nR(1-R), some such simplifying assumptions appear necessary in order to estimate the bias that may result from school nonresponse. Had the sequential search for responding schools continued until two were obtained, the $n\hat{R}_s = (m-1)/(n-1)$ would have been an unbiased estimate of R with $\hat{R}_s(1-\hat{R}_s)/(n-2)$ providing an unbiased estimate for the variance of $\hat{R}_s$. As noted above, truncation destroys this simplicity.

Taking note of the caveat discussed above, the base-year process will be modeled as follows:

(1) The n schools asked to cooperate in a typical stratum will be viewed as a "with replacement" selection of a predetermined number (n) of schools drawn with probabilities P(i);

(2) During the first follow-up a random subsample of s out of the n-m initially uncooperative schools were contacted and student responses were obtained. Subsampling of nonresponding schools must be allowed since noncooperating backup schools were not included in the follow-up.

3.0 Bias Estimation for Base-year Totals

The "biased" school nonresponse base-year estimate for a final stratum total with $m > 0$ can be represented as follows:

$$\hat{Y}(B) = C(n/m) \sum_{i=1}^{n} \rho(i)\, \hat{Y}(i)/nP(i) \qquad (3.1)$$

$$= C\, Y(R)/\hat{R}$$

where

$\rho(i) \equiv$ 1 if sample school-i participated in the base-year and zero otherwise;

$\hat{Y}(i) \equiv$ an estimated total for school-i based on a simple random sample of approximately 18 seniors;

$\hat{Y}(R) \equiv$ an unbiased estimate of the total among the set R of cooperative schools in the stratum;

$\hat{R} \equiv (m/n) = \sum_{i=1}^{n} \rho(i)\, P(i)/nP(i)$ is an unbiased estimate for R the relative size measure for the set R of cooperative schools.

C = a weight smearing factor specific to the final stratum which adjusts for other final strata in the same major stratum which had no cooperating base-year schools; namely, $C(\ell h) = \{1 + [A_o(\ell)/A(\ell h)H(\ell)]\}$ with $A(\ell h)$ denoting the accumulated size measure for final stratum-h in major stratum-ℓ with $A_o(\ell)$ depicting the accumulated size measure for all final strata in major stratum-ℓ with $m(\ell h) = 0$, and with $H(\ell)$ denoting the number of final strata in major stratum-ℓ with $m(\ell h) > 0$.

The follow-up subsample of nonresponding schools can be used to estimate $Y(\bar{R})$ the subpopulation total for the set $\bar{R}$ of uncooperative schools. With a simple random subsample of s (> 0) follow-up responses out of the $n-m$ initially uncooperative schools, the unbiased estimate for $Y(\bar{R})$ is

$$\hat{Y}_F(\bar{R}) = \left(\frac{n-m}{s}\right) \sum_{i=1}^{n} \lambda_F(i)[1-\rho(i)]\, \hat{Y}(i)/nP(i) \qquad (3.2)$$

where

$$\lambda_F(i) = \begin{cases} 1 \text{ if school-}i \text{ initially refuses and is subsequently selected for follow-up} \\ 0 \text{ otherwise.} \end{cases}$$

If $E_F(\ |B)$ denotes expectation over the set of all follow-up subsamples for a given base-year sample, then

$$E_F[\lambda_F(i)|B] = s/(n-m)$$

and

$$E_F[\hat{Y}_F(\bar{R})|B] = Y(\bar{R}) = \sum_{i=1}^{n} [1-\rho(i)]\, \hat{Y}(i)/nP(i) \ . \qquad (3.3)$$

The estimator for $Y(\bar{R})$ in (3.3) is obviously unbiased. When at least one follow-up response is obtained, an unbiased estimate for the population total $Y = Y(R) + Y(\bar{R})$ can be estimated by adding the corresponding statistics from equations (3.1) and (3.2), namely

$$\hat{Y} = \hat{Y}(R) + \hat{Y}_F(\bar{R}) \ . \qquad (3.4)$$

When there are noncooperating base-year schools ($n-m>0$) and no follow-up responses are obtained ($s=0$), follow-up responses from neighboring strata can

be used to approximate $\hat{Y}_F(\bar{R})$. Combining equations (3.1) and (3.4), the school nonresponse bias in $\hat{Y}(B)$ is estimated by

$$\text{bias}\{\hat{Y}(B)\} = C\ \hat{Y}(R)/\hat{R} - \hat{Y}(R) - \hat{Y}_F(\bar{R}) \ . \tag{3.5}$$

The next section will develop the variance formulae for $\hat{\text{bias}}\{\hat{Y}(B)\}$ and suggest an estimator for this variance.

4.0 Variance Approximation for the NLS Bias Estimates

To simplify the variance expressions presented in this section, the within school source of variability will not be shown. Variance estimators will be developed based on a "with replacement" school selection model. These with replacement variance approximations are obtained by substituting estimated school totals $\hat{Y}(i)$ into the appropriate single stage variance estimators. Such estimators include the proper contribution for within school variability and overestimate the between school component due to ignoring the finite population correction at this stage. The single stage analysis that follows will lead, therefore, to single stage variance estimators, which will then be used along with estimated school totals to approximate the variance for the NLS two-stage bias estimate in equation (3.5).

To specify the variance of $\hat{\text{bias}}\{\hat{Y}(B)\}$ conditional expectations and variances will be derived over all possible follow-up subsamples for a given base-year selection of schools. These conditional expectation and variance operations will be depicted by $E_F(\ |B)$ and $\text{Var}_F(\ |B)$. With $E_B(\ )$ and $\text{Var}_B(\ )$ denoting expected values and variances with respect to the base-year school selection, one can write

$$\text{Var}\{\text{bias}[\hat{Y}(B)]\} = \text{Var}_B\ E_F\{\text{bias}[\hat{Y}(B)]|B\} + E_B\ \text{Var}_F\{\text{bias}[\hat{Y}(B)]|B\} \ . \tag{4.1}$$

Recalling equations (3.3) and (3.5) one observes that

$$E_F\{\text{bias}[\hat{Y}(B)]|B\} = C\ \hat{Y}(R)/\hat{R} - \hat{Y}(R) - \hat{Y}(\bar{R}) \tag{4.2}$$

where

$$\hat{Y}(R) = \sum_{i=1}^{n} \rho(i)\, Y(i)/nP(i)$$

$$\hat{Y}(\bar{R}) = \sum_{i=1}^{n} [1-\rho(i)]\, Y(i)/nP(i)$$

$$\hat{R} = \sum_{i=1}^{n} \rho(i)\, P(i)/nP(i) \ .$$

Noting that the quantity in (4.2) is a nonlinear function of three estimated totals and the weight smearing factor C, one can utilize Taylor's linearization technique to approximate its variance. Assuming that C is relatively free of sampling variation compared to the other three statistics, the appropriate linearized value is

$$Z(i) = \{C\rho(i)[RY(i) - P(i)\, Y(\bar{R})]/R^2 - Y(i)\} \ . \tag{4.3}$$

The first term in (4.1) is therefore

$$\mathrm{Var}_B E_F\{\mathrm{bias}[\hat{Y}(B)|B\} = \sum_{i=1}^{N} P(i) \left[\frac{Z(i)}{P(i)} - Z(+)\right]^2 /n \tag{4.4}$$

where

$$Z(+) = \sum_{i=1}^{N} Z(i) = \sum_{i=1}^{N} Y(i) = -Y \ .$$

Letting

$$P_R(i) = P(i)/R = A(i)/A(R)$$

denote the conditional probability of selecting a school-i on a specific draw given that school-i belongs to the set R of cooperative schools, the expression in equation (4.4) becomes with substitution for Z(i) from (4.3)

$$\sum_{i=1}^{N} P(i) \left\{ C\,\rho(i) \left[\frac{Y(i)}{P_R(i)} - Y(R)\right] /R^2 - \left[\frac{Y(i)}{P(i)} - Y\right] \right\}^2 /n$$

$$= C^2 \sum_{i\in R} P_R(i) \left[\frac{Y(i)}{P_R(i)} - Y(R)\right]^2 /E(m)R^2$$

$$+ \sum_{i=1}^{N} P(i) \left[\frac{Y(i)}{P(i)} - Y\right]^2 /n$$

$$- 2C \sum_{i\in R} P_R(i) \left[\frac{Y(i)}{P_R(i)} - Y(R)\right] \left[\frac{Y(i)}{P(i)} - Y\right] /E(m) \ . \tag{4.5}$$

Recasting (4.2) as $C\ \hat{Y}(R)/\hat{R} - \hat{Y}$ where $\hat{Y} = \hat{Y}(R) + \hat{Y}(\bar{R})$ is the unbiased estimate for the NLS population total Y, the variance expression in (4.5) gains some intuitive appeal. The first term in (4.5) would be the variance of $C\ \hat{Y}(R)/\hat{R}$ if a fixed sample of $E(m) = nR$ schools was selected directly from the subpopulation R of cooperating schools with replacement and with probabilities $P_R(i)$ where $\hat{R} = R$ and C are known without error. The second term in (4.5) is the familiar PPS with replacement variance formulae for the unbiased total estimate $\hat{Y}$. The third term in (4.5) is minus twice a reasonable looking covariance between $C\ \hat{Y}(R)/\hat{R}$ and $\hat{Y}$.

To evaluate the second term in (4.1) which arises due to subsampling of nonresponding schools for follow-up, observe that

$$\mathrm{Var}_F\{\mathrm{bias}[\hat{Y}(B)]\,|\,B\} = \mathrm{Var}_F\{\hat{Y}_F(\bar{R})\,|\,B\} .$$

Recalling the form of $\hat{Y}_F(\bar{R})$ from equation (3.2), $\hat{Y}_F(\bar{R})$ can be recast as

$$\hat{Y}_F(\bar{R}) = (1-\hat{R}) \sum_{i=1}^{n} \lambda_F(i)[1-\rho(i)]\ Y(i)/sP(i) . \qquad (4.6)$$

The sum following $(1-\hat{R})$ in (4.6) is the average of $\hat{Y}(i)/P(i)$ over the s initially nonresponding schools which participate in the follow-up. Viewing the $k = 1(1)(n-m)$ initial nonrespondents as a population from which s members were selected via simple random sampling with $y(k) = Y(k)/P(k)$ the observed variate value for sample school-k, then the sum in (4.6) can be written as

$$\bar{y}_F(\bar{R}) = \sum_{k=1}^{s} y(k)/s$$

and therefore

$$E_F\{\bar{y}_F(\bar{R})\,|\,B\} = \bar{y}(\bar{R}) = \sum_{k=1}^{n-m} y(k)/(n-m) = \sum_{k=1}^{n-m} Y(k)/(n-m)P(i)$$

$$= \sum_{i=1}^{n} [1-\rho(i)]\ \hat{Y}(i)/(n-m)P(i) .$$

Similarly,

$$\mathrm{Var}_F\{\bar{y}_F(\bar{R})|B\} = \left[1 - \frac{s}{(n-m)}\right] \sum_{k=1}^{n-m} [y(k) - \bar{y}(\bar{R})]^2/s(n-m-1)$$

$$= \left(\frac{n-m-s}{n-m}\right) \sum_{k=1}^{n-m-1} \sum_{\ell=k+1}^{n-m} [y(k) - y(\ell)]^2/s(n-m)(n-m-1). \tag{4.7}$$

Extending the double sum in (4.7) over all $n(n-1)/2$ pairs of inititally contacted schools yields

$$\mathrm{Var}_F\{\bar{y}_F(\bar{R})|B\} = \left(\frac{n-m-s}{n-m-1}\right) \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} [1-\rho(i)][1-\rho(j)][y(i)-y(j)]^2/2sn^2(1-\hat{R})^2.$$

Since $\hat{Y}_F(\bar{R}) = (1-\hat{R})\bar{y}_F(\bar{R})$, the conditional variance of $\hat{Y}_F(\bar{R})$ for a given base year school selection has the form above with $(1-\hat{R})^2$ deleted from the denominator. With s representing the number of nonresponding primary schools from a typical stratum, the expected value of s over all possible base year selections is $E_B(s) = 2(1-R)$. The expected value of the double sum of squared differences in $\mathrm{Var}_F\{\bar{y}_F(\bar{R})|B\}$ above is

$$n(n-1) \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} P(i)\,P(j)[1-\rho(i)][1-\rho(j)][y(i) - y(j)]^2.$$

Using $P_{\bar{R}}(i) = P(i)/(1-R)$ to denote the relative sizes among the population $\bar{R}$ of noncooperative schools, the quantity above can be recast as

$$n(n-1) \sum_{i \in \bar{R}} P_{\bar{R}}(i) \left[\frac{Y(i)}{P_{\bar{R}}(i)} - Y(\bar{R})\right]^2 .$$

Combining these results, the second term in the variance of bias $[\hat{Y}(B)]$ is approximately

$$E_B \mathrm{Var}_F\{Y_F(\bar{R}) \mid B\} \doteq \left[\frac{(n-1)(n-2)}{2n(n-nR-1)}\right] \sum_{i \in \bar{R}} P_{\bar{R}}(i) \left[\frac{Y(i)}{P_{\bar{R}}(i)} - Y(\bar{R})\right]^2 \qquad (4.8)$$

Putting the expressions in (4.5) and (4.8) together, one obtains

$$\mathrm{Var}\{\mathrm{bias}[\hat{Y}(B)]\} \doteq C^2 \sigma_R^2/E(m)R^2 + \sigma_\tau^2/n - 2C\sigma_{R\tau}/E(m)$$

$$+\left[\frac{(n-1)(n-2)}{2(n-nR-1)}\right]\sigma_{\bar{R}}^2/n \qquad (4.9)$$

where σ_τ^2 is the familiar PPS with replacement variance expression for the unbiased stratum total estimate. The quantities σ_R^2 and $\sigma_{\bar{R}}^2$ represent similar PPS with replacement variances for fixed sized samples drawn with probabilities $P_R(i) = P(i)/R$ and $P_{\bar{R}}(i) = P(i)/(1-R)$ respectively from the subpopulations of responding R and nonresponding $\bar{R}$ schools. The $\sigma_{R\tau}$ denotes the covariance type component from equation (4.5). Recalling the definition of the covariance component in equation (4.5), one can show that $\sigma_{R\tau} = \sigma_R^2/R$ which leads to

$$\mathrm{Var}\{\mathrm{bias}[\hat{Y}(B)]\} \doteq (C^2-2RC)\sigma_R^2/E(m)R^2 + \sigma_\tau^2/n$$

$$+ \left[\frac{(n-1)(n-2)}{2(n-nR-1)}\right]\sigma_{\bar{R}}^2/n \; . \qquad (4.10)$$

An alternative form for equation (4.10) can be obtained by observing that

$$\sigma_\tau^2 = \sigma_R^2/R + \sigma_{\bar{R}}^2/(1-R) + R(1-R)\left[\frac{Y(R)}{R} - \frac{Y(\bar{R})}{(1-R)}\right]^2 .$$

The equality above leads to

$$\mathrm{Var}\{\mathrm{bias}[\hat{Y}(B)]\} \doteq \left(\frac{C-R}{R}\right)^2 \sigma_R^2/E(m) + \sigma_R^2/E(n-m) + R(1-R)\left[\frac{Y(R)}{R} - \frac{Y(\bar{R})}{(1-R)}\right]^2/n$$

$$+ \left[\frac{(n-1)(n-2)}{2(n-nR-1)}\right]\sigma_{\bar{R}}^2/n \; . \qquad (4.11)$$

5.0 Variance Estimation for the NLS Bias Measure

To estimate the variance expression in equation (4.11), an approximation for the linearized value Z(i) presented in equation (4.3) can be used. Letting

$$\hat{Z}(i) = \{C\rho(i)\ [\hat{R}\hat{Y}(i) - P(i)\ \hat{Y}(R)]/\hat{R}^2 - \hat{Y}(i)\} \tag{5.1}$$

and defining

$$\begin{aligned} z(i) &= \hat{Z}(i)/P(i) \\ &= \{C\rho(i)\ [y(i) - \bar{y}_r]/\hat{R} - y(i)\} \end{aligned} \tag{5.2}$$

where

$$y(i) = \hat{Y}(i)/P(i)$$

and

$$\bar{y}_r = \hat{Y}(R)/\hat{R} = \sum_{i=1}^{n} \rho(i)Y(i)/m\ P(i)$$

then

$$s_z^2/n = \sum_{i=1}^{n} [z(i) - \bar{z}]^2/n(n-1) \tag{5.3}$$

estimates the variance of bias$[\hat{Y}(B)]$ when all (n-m) nonresponding schools are followed up; that is, when there is no subsampling of noncooperating schools. When there is subsampling, then $\hat{Y}(i)$ is not available for all n schools initially asked to participate. In this situation it is necessary to estimate (5.3) based on the follow-up subsample results, and to add an additional term to include the subsampling variability. To see how (5.3) can be estimated, it is helpful to expand s_z^2/n using the definition of z(i) as follows

$$s_z^2/n = \sum_{i=1}^{n} \{C\rho(i)[y(i) - \bar{y}_r]/\hat{R} - [y(i) - \bar{y}]\}^2/n(n-1)$$

$$= C^2 \sum_{i=1}^{n} \rho(i)[y(i) - \bar{y}_r]^2/n(n-1)\hat{R}^2$$

$$- 2C \sum_{i=1}^{n} \rho(i)[y(i) - \bar{y}_r][y(i) - \bar{y}]/n(n-1)\hat{R}$$

$$+ \sum_{i=1}^{n} [y(i) - \bar{y}]^2/n(n-1) \quad . \tag{5.4}$$

The first two terms in (5.4) are related to the mean square of y(i) variates among the m initially responding schools, say

$$\delta_r^2 = \sum_{i=1}^{n} \rho(i)[y(i) - \bar{y}_r]^2/(m-1)$$

such that

$$\delta_z^2 = C^2(m-1)\delta_r^2/m(n-1)\hat{R} - 2C(m-1)\delta_r^2/m(n-1) + \delta_y^2/n$$

$$= (C^2 - 2\hat{R}C)(m-1)\delta_r^2/m(n-1)\hat{R} + \delta_y^2/n \quad . \tag{5.5}$$

The first term in (5.5) can be estimated directly from the base-year responding schools. For strata with m = 1 responding school, the first term in (5.5) will drop out since $(m-1)\delta_r^2 = 0$. When m = 0, no such term appears in the variance estimator. The second term in (5.5) can be expanded further as follows:

$$\delta_y^2/n = \sum_{i=1}^{n} \sum_{j \neq i} \{\rho(i) + [1-\rho(i)]\}\{\rho(j) + [1-\rho(j)]\}[y(i) - y(j)]^2/2n^2(n-1) \quad .$$

Exploiting this identity one obtains

$$\delta_y^2/n = \sum_{i=1}^{n} \sum_{j \neq i} \rho(i)\rho(j)[y(i) - y(j)]^2/2n^2(n-1)$$

$$+ \sum_{i=1}^{n} \rho(i) \sum_{j \neq i} [1 - \rho(j)][y(i) - y(j)]^2/n^2(n-1)$$

$$+ \sum_{i=1}^{n} \sum_{j \neq i} [1 - \rho(i)][1 - \rho(j)][y(i) - y(j)]^2/2n^2(n-1). \tag{5.6}$$

The first term in (5.6) is clearly

$$\hat{R}^2(m-1)s_r^2/m(n-1) \ . \tag{5.7}$$

The second term in (5.6) can be recast as

$$\sum_{i=1}^{n} \rho(i) \sum_{j=1}^{n} [1-\rho(j)]\,[y(i) - y(j)]^2/n^2(n-1) \ . \tag{5.8}$$

Recalling the subsampling indicators $\lambda_F(i)$ with $E_F\{\lambda_F(i)|B\} = s/(n-m)$, an unbiased estimate for (5.8) is

$$(n-m) \sum_{i=1}^{n} \rho(i) \sum_{j=1}^{n} \lambda_F(j)\,[1 - \rho(j)]\,[y(i) - y(j)]^2/sn^2(n-1) \ . \tag{5.9}$$

Letting

$$\bar{y}_r = \sum_{j=1}^{n} \lambda_F(j)\,[1 - \rho(j)]y(j)/s \quad \text{and}$$

$$s_r^2 = \sum_{j=1}^{n} \lambda_F(j)[1 - \rho(j)][y(j) - \bar{y}_r]^2/(s-1), \text{ the}$$

second term estimator in (5.9) becomes

$$(1-\hat{R}) \sum_{i=1}^{n} \rho(i) \sum_{j=1}^{n} \lambda_F(j)\,[1- (j)]\{[y(i)-\bar{y}_r] - [y(j)-\bar{y}_r] + [\bar{y}_r-\bar{y}_r]\}^2/sn(n-1)$$

$$= \hat{R}(1-\hat{R})(m-1)s_r^2/m(n-1) + \hat{R}(1-\hat{R})(s-1)s_r^2/s(n-1) + \hat{R}(1-\hat{R})[\bar{y}_r - \bar{y}_r]^2/(n-1). \tag{5.10}$$

Noting that

$$s_r^2 = \sum_{i=1}^{n} \sum_{j\neq i} \lambda_F(i)\,\lambda_F(j)[1 - \rho(i)][1 - \rho(j)][y(i) - y(j)]^2/2s(s-1),$$

the third term in (5.6) can be estimated unbiasedly by

$$(1-\hat{R})(n-m-1)s_r^2/n(n-1) \ . \tag{5.11}$$

Combining (5.7), (5.10), and (5.11) the unbiased estimate for δ_y^2/n is

$$\hat{\delta}_y^2/n = R(m-1)\delta_r^2/m(n-1) + (1-\hat{R}((ns-m-s)\delta_{\bar{r}}^2/sn(n-1) + \hat{R}(1-\hat{R})(\bar{y}_r - \bar{y}_{\bar{r}})^2/(n-1) \quad . \tag{5.12}$$

Substituting this estimate for δ_y^2/n in (5.5) one obtains

$$\hat{\delta}_z^2/n = (\hat{C}-\hat{R})^2(m-1)\delta_r^2/m(n-1)\hat{R} + (1-\hat{R})(ns-m-s)\delta_{\bar{r}}^2/sn(n-1) + \hat{R}(1-\hat{R})(\bar{y}_r - \bar{y}_{\bar{r}})^2/(n-1) \quad . \tag{5.13}$$

To estimate the subsampling component of variation in bias$[\hat{Y}(B)]$, the relations

$$\mathrm{Var}_F\{\mathrm{bias}[\hat{Y}(B)]|B\} = \mathrm{Var}_F\{\hat{Y}_F(\bar{R})|B\} = (1-\hat{R})^2\mathrm{Var}_F\{\bar{y}_F(\bar{R})|B\}$$

are useful, where in the shorthand notation used in (5.13) $\bar{y}_r(\bar{R}) = \bar{y}_{\bar{r}}$. Recalling that the s follow-up schools represent a simple random sample from the (n-m) initially uncooperative schools, one notes that

$$\mathrm{var}_F\{\bar{y}_{\bar{r}}|B\} = \left[1 - \frac{s}{(n-m)}\right]\delta_{\bar{r}}^2/s \tag{5.14}$$

is an unbiased estimate for $\mathrm{Var}_F\{\bar{y}_F(\bar{R})|B\}$ and in turn for $E_B\mathrm{Var}_F\{\bar{y}_F(\bar{R})|B\}$. Therefore

$$E_B\mathrm{Var}_F\{\mathrm{bias}[\hat{Y}(B)]|B\} = (1-\hat{R})(n-m-s)\delta_{\bar{r}}^2/sn \quad . \tag{5.15}$$

Adding (5.15) to (5.13), the required variance estimator is

$$\text{var}\{\text{bias}[\hat{Y}(B)]\} = \{(C-\hat{R})^2(m-1)\delta_r^2/n + (1-\hat{R})(n-m-1)\delta_{\bar{r}}^2/s + \hat{R}(1-\hat{R})(\bar{y}_r-\bar{y}_{\bar{r}})^2\}/(n-1) \quad . \tag{5.16}$$

In the following section, these results will be extended to provide bias estimates for ratio means and proportions as well as variance estimates for these bias measures.

6.0 Bias Analysis for NLS Ratio Means and Proportions

The school nonresponse biased base year estimate for an NLS ratio mean or proportion can be represented by:

$$\hat{\theta}(B) = \sum_{h=1}^{H} C(h)\hat{Y}_R(h)/\hat{R}(h) \Big/ \sum_{h=1}^{H} C(h)\hat{X}_R(h)/\hat{R}(h) \tag{6.1}$$

where

$$\hat{X}_R(h) = \sum_{i=1}^{n(h)} \rho(hi)\hat{X}(hi)/n(h)P(hi)$$

with $\hat{X}(hi)$ denoting, for example, the estimated number of black seniors in sample school-i of final stratum-h. The $\hat{Y}_R(h)$ total for stratum-h is defined similarly with $\hat{Y}(hi)$ depicting, for example, the estimated number of black seniors in school-hi who would respond to an NLS question in a particular way. For this example $\hat{\theta}(B)$ represents the biased estimate for the national proportion of black seniors who would respond to a particular NLS question in a specific way. Letting

$$\hat{Y}_B(h) \equiv C(h)\hat{Y}_R(h)/\hat{R}(h) \text{ and } \hat{X}_B(h) = C(h)\hat{X}_R(h)/\hat{R}(h) \quad ,$$

$$\hat{\theta}(B) = \sum_{h=1}^{H} \hat{Y}_B(h) \Big/ \sum_{h=1}^{H} \hat{X}_B(h) = \hat{Y}(B)/\hat{X}(B) \quad . \tag{6.2}$$

An unbiased estimate for θ can be formed using the s(h) follow-up responses obtained from a subsample of the nonresponding schools. With $\hat{Y}_{F\bar{R}}(h)$ and $\hat{X}_{F\bar{R}}(h)$ representing estimated totals for the stratum-h population of uncooperative schools where

$$\hat{X}_{F\bar{R}}(h) = [1 - \hat{R}(h)] \sum_{i=1}^{n(h)} \lambda_F(hi)\,[1 - \rho(hi)]\hat{X}(hi)/s(h)P(hi)$$

$$= [1 - \hat{R}(h)]\bar{x}_r(h)$$

and $\hat{Y}_{F\bar{R}}(h)$ has an analogous form, a reasonably unbiased estimate for θ is

$$\hat{\theta} = \sum_{h=1}^{H} [\hat{Y}_R(h) + \hat{Y}_{F\bar{R}}(h)] / \sum_{h=1}^{H} [\hat{X}_R(h) + \hat{X}_{F\bar{R}}(h)]$$

$$= \sum_{h=1}^{H} \hat{Y}(h) / \sum_{h=1}^{H} \hat{X}(h)$$

$$= \hat{Y}/\hat{X} . \qquad (6.3)$$

Combining the estimators in (6.2) and (6.3) the school nonresponse bias in $\hat{\theta}(B)$ is estimated by

$$\text{bias}[\hat{\theta}(B)] = \hat{\theta}(B) - \hat{\theta} = \frac{\hat{Y}(B)}{\hat{X}(B)} - \frac{\hat{Y}}{\hat{X}} \qquad (6.4)$$

To determine a variance estimator for the bias measure in (6.4), a separate linearization can be formed for the two ratios $\hat{\theta}(B)$ and $\hat{\theta}$. The combined linearization for bias$[\hat{\theta}(B)]$ will be the difference between the two separate linearizations. Taking partial derivatives of $\hat{\theta}(B)$ with respect to $\hat{Y}_R(h)$, $\hat{X}_R(h)$, and $\hat{R}(h)$ one obtains

$$L_B(hi) = C(h)\rho(hi)\{[\hat{Y}(hi) - P(hi)\hat{Y}_B(h)]$$

$$-\hat{\theta}(B)[\hat{X}(hi) - P(hi)\hat{X}_B(h)]\}/\hat{R}(h)\hat{X}(B) .$$

Letting

$$\ell_B(hi) = L_B(hi)/P(hi)$$

$$= C(h)\rho(hi)\{[y(hi)-\bar{y}_r(h)] - \hat{\theta}(B)[x(hi)-x_r(h)]\}/\hat{R}(h)\hat{X}(B)$$

where

$$y(hi) = \hat{Y}(hi)/P(hi) \quad \text{and}$$

$$\bar{y}_r(h) = \sum_{i=1}^{n(h)} \rho(hi)\hat{Y}(hi)/m(h)P(h)$$

with $\bar{x}_r(h)$ defined similarly, one can form

$$t(hi) = [y(hi) - \hat{\theta}(B)x(hi)]/\hat{X}(B)$$

which leads to

$$\ell_B(hi) = C(h)\,\rho(hi)[t(hi)-\bar{t}_r(h)]/R(h) \quad . \tag{6.5}$$

Taking partial derivatives of $\hat{\theta}$ with respect to $\hat{Y}_R(h)$, $\hat{Y}_{F\bar{R}}(h)$, $\hat{X}_R(h)$ and $\hat{X}_{F\bar{R}}(h)$ one obtains

$$L(hi) = [\hat{Y}(hi) - \hat{\theta}\hat{X}(hi)]/\hat{X}$$

and

$$\ell(hi) = [y(hi) - \hat{\theta}x(hi)]/\hat{X} \quad . \tag{6.6}$$

Subtracting (6.6) from (6.5) yields

$$v(hi) = \ell_B(hi) - \ell(hi) = C(h)\,\rho(hi)[t(hi)-t_r(h)]/R(h) - \ell(hi) \ ,$$

the linearized value for estimating $\mathrm{Var}_B E_F\{\mathrm{bias}[\hat{\theta}(B)]|B\}$. Using the expansion developed in (5.4) one finds that

$$s_v^2(h)/N(h) = C^2(h)[m(h)-1]\, s_{tr}^2(h)/m(h)[n(h)-1]\hat{R}(h)$$

$$- 2C(h)[m(h)-1]\, s_{t\ell_r}^2(h)/m(h)[n(h)-1]$$

$$+ s_\ell^2(h)/n(h) \tag{6.7}$$

where

$$s_{tr}^2 = \sum_{i=1}^{n(h)} p(hi)[t(hi) - \bar{t}_r(h)]^2/[m(h) - 1]$$

$$s_{t\ell r}^2 = \sum_{i=1}^{n(h)} p(hi)[t(hi) - \bar{t}_r(h)][\ell(hi) - \ell_r(h)]/[m(h) - 1]$$

and

$$s_\ell^2(h) = \sum_{i=1}^{n(h)} [\ell(hi) - \bar{\ell}(h)]^2/[n(h) - 1].$$

Recalling the expansion for s_y^2/n developed in (5.6) one notes that $s_\ell^2(h)/n(h)$ can be estimated by

$$\hat{s}_\ell^2(h)/n(h) = \hat{R}(h)[m(h) - 1]\, s_{\ell r}^2(h)/m(h)[n(h) - 1]$$
$$+ [1 - \hat{R}(h)][n(h)s(h) - m(h) - s(h)]s_{\ell\bar{r}}^2(h)/s(h)n(h)[n(h)-1]$$
$$+ \hat{R}(h)[1 - \hat{R}(h)][\bar{\ell}_r(h) - \bar{\ell}_{\bar{r}}(h)]^2/[n(h) - 1] \quad . \tag{6.8}$$

Now, equations (6.7) and (6.8) will be combined and simplified. Defining

$$w(hi) = [C(h)\, t(hi) - \hat{R}(h)\ell(hi)] \ ,$$

the sum of equations (6.7) and (6.8) is

$$\hat{s}_v(h)/n(h) = [m(h) - 1]s_{wr}^2(h)/m(h)[n(h) - 1]\hat{R}(h)$$
$$+ [1 - \hat{R}(h)][n(h)s(h) - m(h) - s(h)]s_{\ell\bar{r}}^2(h)/s(h)n(h)[n(h) - 1]$$
$$+ \hat{R}(h)[1 - \hat{R}(h)][\bar{\ell}_r(h) - \bar{\ell}_{\bar{r}}(h)]^2/[n(h) - 1] \quad . \tag{6.9}$$

To complete the variance estimator for bias$[\hat{\theta}(B)]$ one obtains by analogy with (5.15)

$$E_B Var_F\{\text{bias}[\hat{\theta}(B)]|B\} = [1 - \hat{R}(h)][n(h) - m(h) - s(h)]s_{\ell\bar{r}}^2(h)/n(h)s(h), \tag{6.10}$$

an estimate for the subsampling contribution to Var{bias[$\hat{\theta}(B)$]}. Finally, combining (6.9) with (6.10), the required variance estimator is obtained, namely

$$\text{var}\{\text{bias}[\hat{\theta}(B)]\} = \sum_{h=1}^{H} \{[m(h) - 1]\, s^2_{wr}(h)/m(h)[n(h) - 1]\, \hat{R}(h)$$

$$+ [1 - \hat{R}(h)]\, [n(h) - m(h) - 1] s^2_{\ell r}(h)/s(h)[n(h) - 1]$$

$$+ \hat{R}(h)[1 - \hat{R}(h)]\, [\bar{\ell}_r(h) - \bar{\ell}_r(h)]^2/[n(h) - 1]\} \qquad (6.11)$$

where

$$w(hi) = C(h)[y(hi) - \hat{\theta}(B)\hat{x}(hi)]/\hat{X}(B) - \hat{R}(h)[y(hi) - \hat{\theta}x(hi)]/\hat{X}$$

$$\ell(hi) = [y(hi) - \hat{\theta}x(hi)]/\hat{X}$$

$$s^2_{wr}(h) = \sum_{i=1}^{n(h)} \rho(hi)[w(hi) - \bar{w}_r(h)]^2/[m(h) - 1]$$

with

$$\bar{w}_r(h) = \sum_{i=1}^{n(h)} \rho(hi)w(hi)/m(h)$$

and

$$s^2_{\ell r}(h) = \sum_{i=1}^{n(h)} \lambda_F(hi)[1 - \rho(hi)[\ell(hi) - \bar{\ell}_r(h)]^2/[s(h) - 1]$$

with

$$\bar{\ell}_r(h) = \sum_{i=1}^{n(h)} \lambda_F(hi)[1 - \rho(hi)]\ell(hi)/s(h).$$

The $\bar{\ell}_r(h)$ quantity in (6.11) represents the simple average of $\ell(hi)$ variates over the m(h) initially responding schools in stratum-h; namely

$$\bar{\ell}_r(h) = \sum_{i=1}^{n(h)} \rho(hi)\ell(hi)/m(h).$$

Examining the variance estimator in (6.11) it is clear that the first term involving $[m(h)-1]s^2_{wr}(h)$ drops out when m(h) = 0 or 1. When s(h) [the number of follow-up schools] is less than [n(h) - m(h)], the number of initially uncooperative schools contacted in stratum-h, then s(h) must be two or greater to provide a subsampling variance component estimate $s^2_{\bar{\ell}r}(h)$. If s(h) = 1, stratum-h should be collapsed with a neighboring stratum-h' so that s(h) + s(h') $\geq$ 2 and a pooled estimate $s^2_{\bar{\ell}r}(h+h')$ can be produced. When [n(h)-m(h)] > 0 and s(h) = 0 both $\bar{\ell}_r(h)$ and $s^2_{\bar{\ell}r}(h)$ must be borrowed fro a neighboring stratum.

7.0 Testing for School Nonresponse Bias

With the variance estimator proposed in equation (6.11), a normal theory test for significant school nonresponse bias can be performed. Two additional summary statistics which may be of interest are the relative bias

$$\text{rel-bias } [\hat{\theta}(B)] = \text{bias}[\hat{\theta}(B)]/\hat{\theta}$$

and the "so-called" bias ratio

$$\text{bias-ratio } [\hat{\theta}(B)] = \text{bias } [\hat{\theta}(B)]/\text{var}[\hat{\theta}(B)] \quad .$$

The bias-ratio can be used to determine the impact of bias on the probability that a confidence interval of the form

$$\hat{\theta}(B) \pm K \text{ var}[\hat{\theta}(B)]$$

will contain the true population value θ. The relative bias and bias-ratio measures can be averaged over similar statistics to provide summary bias indicators.